

La Base FRANTEXT

Documentation de l'utilisateur

documentation réunie par Josette Lecomte – Ingénieure de recherche

décembre 2002



Analyse et Traitement Informatique de la Langue Française
44, avenue de la Libération - BP 30687 - 54063 Nancy Cedex
Téléphone : 03 83 96 21 76 Télécopie : 03 83 97 24 56

www.atilf.fr contact@atilf.fr



Documentation de l’utilisateur / version décembre 2002
(documentation réunie par Josette Lecomte, Ingénieure de Recherche)

La Base Frantext

Conception et réalisation informatiques : Jacques Dendien

La base Frantext (1992) peut être définie comme l’association d’une part d’un vaste corpus de textes littéraires français, et d’autre part d’un logiciel offrant une interface Web avec des possibilités d’interrogation de consultation et d’hyper-navigation. Historiquement, le but premier de ce corpus textuel était de permettre la constitution d’une base d’exemples destinée aux rédacteurs des articles du TLF.
La base Frantext contient 3737 textes (au 1er janvier 2004) appartenant aux domaines des sciences, des arts, de la littérature, des techniques, qui couvrent 5 siècles de littérature (du XVIe au XXe siècle). Elle est accessible sur Internet, moyennant un abonnement.
Deux versions de Frantext sont proposées :
- L’intégralité de la base (3737 textes, environ 210 millions d’occurrences, environ un millier d’auteurs). Les œuvres se répartissent pour 80% d’œuvres littéraires et 20% d’œuvres scientifiques ou techniques. Il est possible d’effectuer des recherches à différents niveaux : simples ou complexes.
- Une sous-partie constituée de 1940 œuvres en prose des XIXe et XXe siècles, soit environ 127 millions d’occurrences, qui ont fait l’objet d’un codage grammatical selon les Parties du Discours. Aux fonctionnalités du Frantext intégral, ont été ajoutées des possibilités de requêtes portant sur les codes grammaticaux.

L'accès à la base Frantext est réservé aux abonnés.

Conditions d'abonnement

Tout organisme d'enseignement ou de recherche, bibliothèque... a la possibilité de s'abonner à la base de données textuelles **FRANTEXT** du laboratoire ATILF.
Les abonnés à la base **FRANTEXT** bénéficient de l'accès libre à la ressource "**Encyclopédie de Diderot et d'Alembert**".
Le coût de l’abonnement fixé à 310€ H.T. (prix 2004) à l’année permet de connecter 50 machines simultanément.

Contacts :

Pour le contenu textuel de la base :	Service des Bases textuelles (**sbt@atilf.fr**)
Pour les aspects informatiques :	Jacques Dendien (**jacques.dendien@atilf.fr**)
Pour le maniement du logiciel :	Pascale Bernard (**pascale.bernard@atilf.fr**)
Pour l'abonnement :	**frantext@atilf.fr**

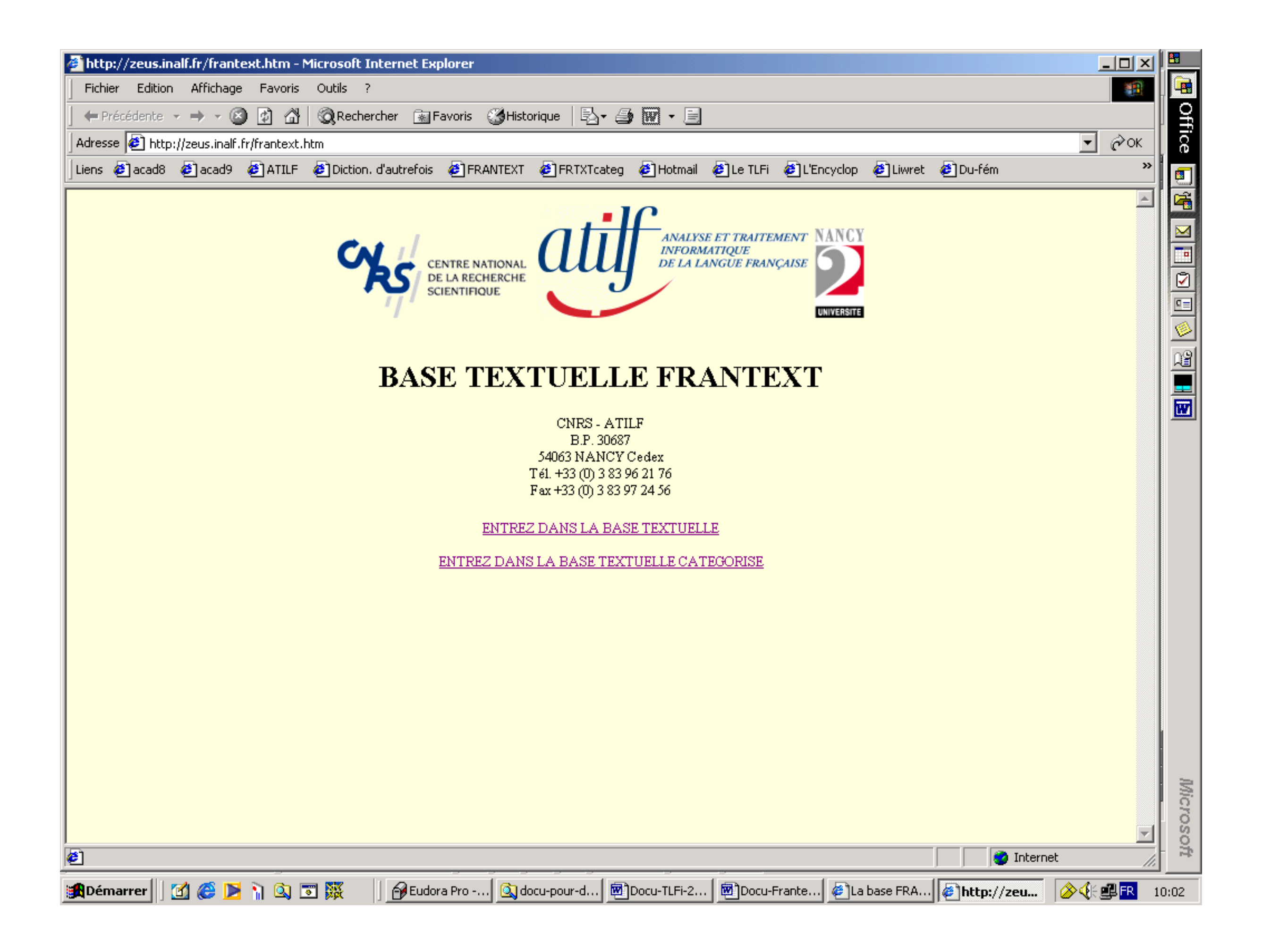


ATILF (UMR 7118 CNRS - Université Nancy 2)
44, avenue de la Libération - BP 30687 - 54063 Nancy Cedex
Tél. 03 83 96 21 76 Fax 03 83 97 24 56
Site : www.atilf.fr Courriel : contact@atilf.fr





Écran d'accueil sur les sites FRANTEXT

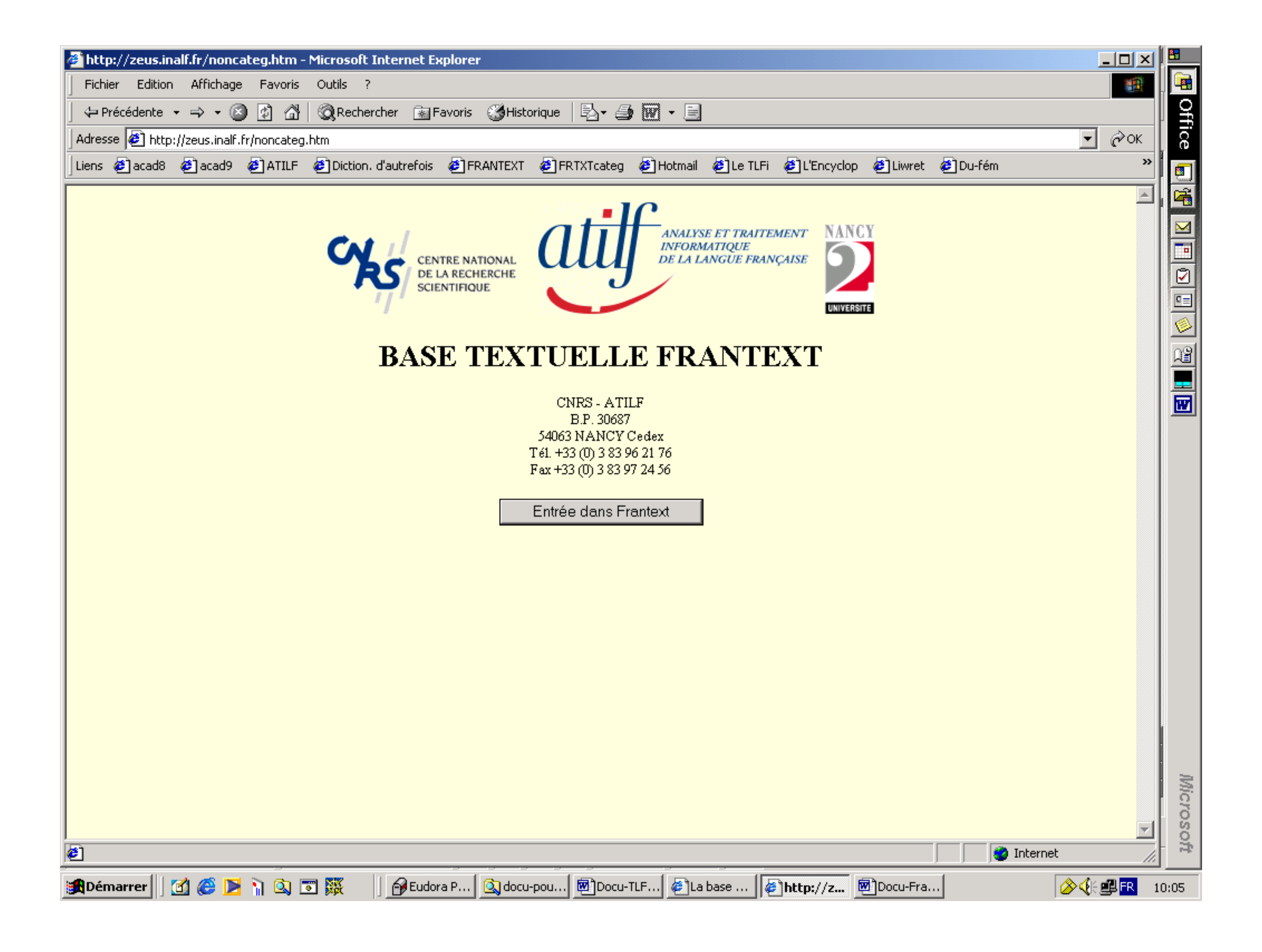


<u>Avertissement :</u>

La documentation qui suit est le reflet de la documentation en ligne, figé à un moment donné. Elle ne se substitue en aucun cas à cette doc. en ligne, que Jacques Dendien fait évoluer régulièrement. Elle ne dispense donc pas d'aller consulter cette aide en ligne lors des requêtes sur la base Frantext.

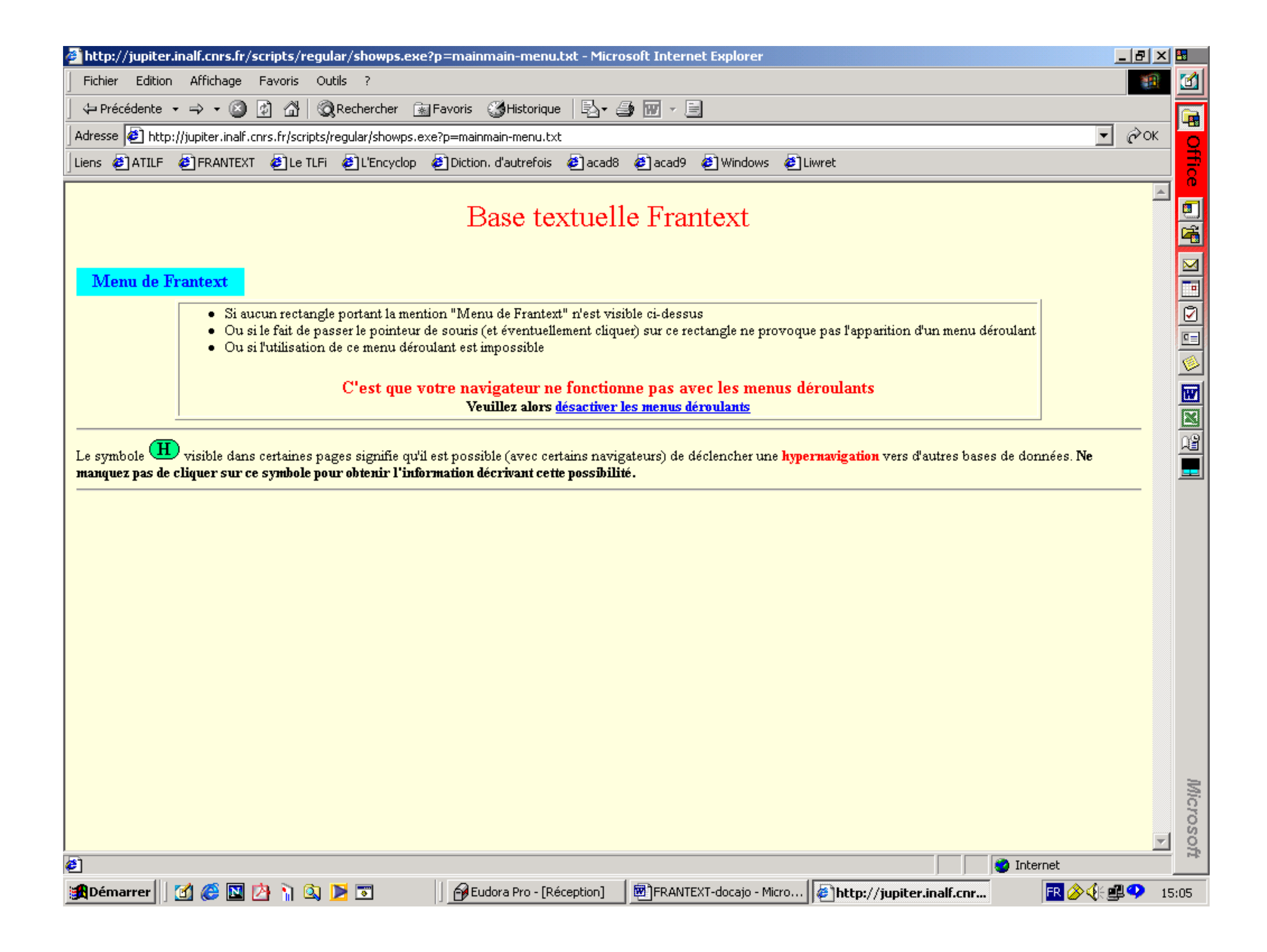
1ère section : Frantext non catégorisé

Écran d'accueil sur le site FRANTEXT non catégorisé

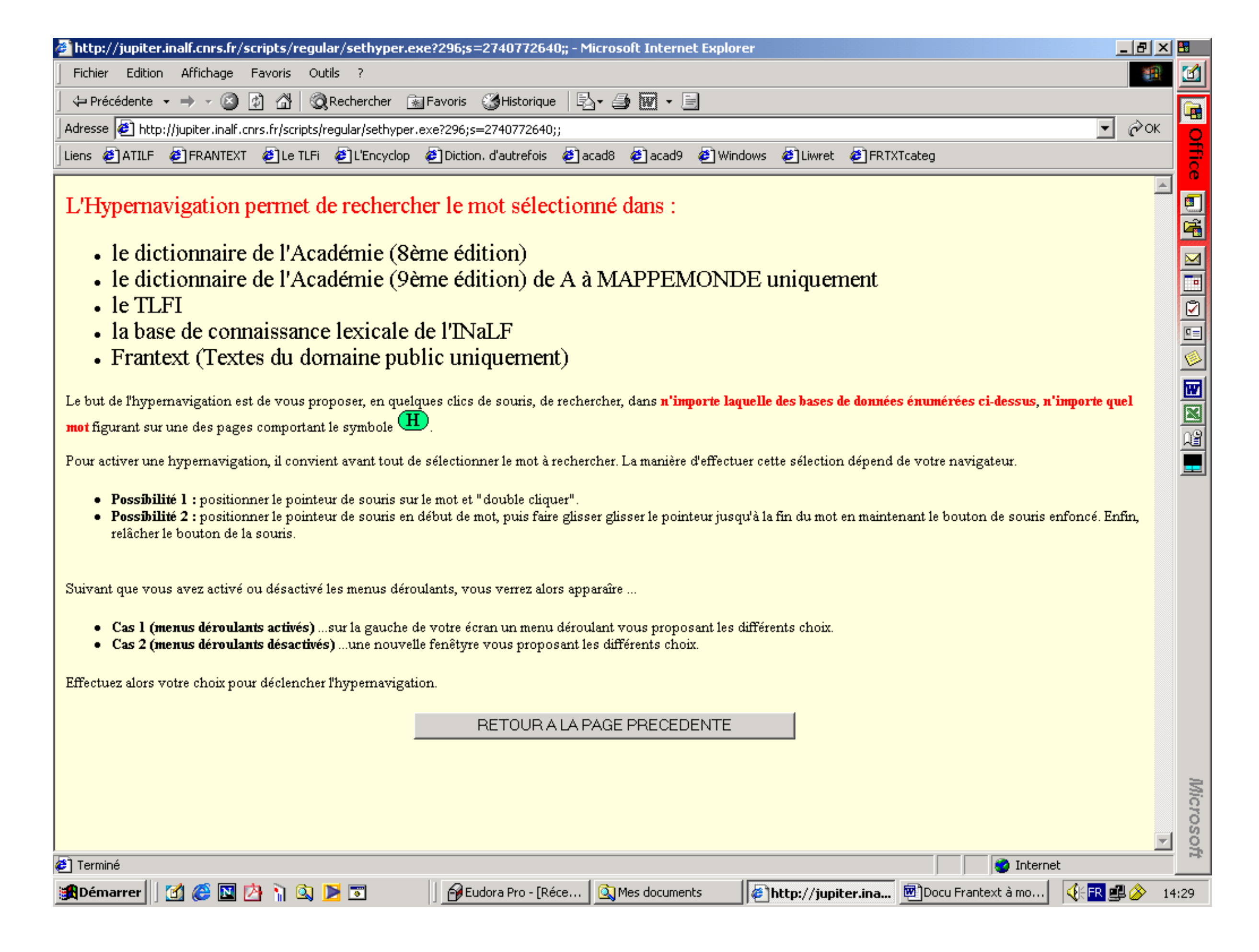


Entrée dans FRANTEXT non catégorisé

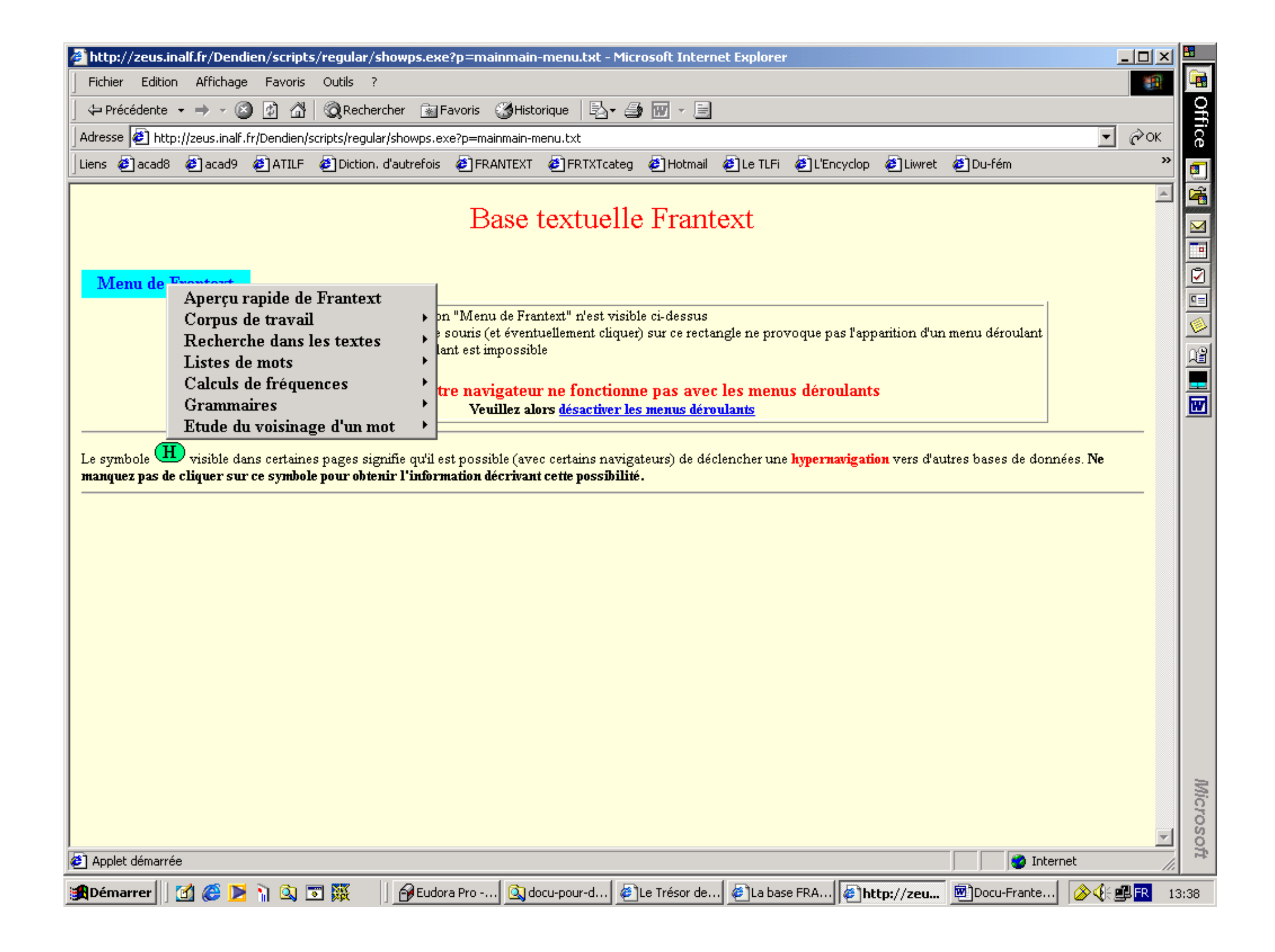
Présentation :



Quelques détails sur l'hypernavigation :



Menu de FRANTEXT



Présentation rapide de la Base textuelle FRANTEXT

1. Avertissement pour les utilisateurs non familiarisés avec WWW.

L'association des services de FRANTEXT et des services offerts par les logiciels de consultation de World Wide Web procure à l'utilisateur un confort d'utilisation appréciable.
Les fonctions intégrées dans les navigateurs Internet (sauvegarde, marquage des pages, gestion de l'historique des pages, etc.) sont autant de services qui viennent se surajouter aux services de FRANTEXT proprement dits.
Pour tirer le meilleur parti de FRANTEXT, l'utilisateur est censé connaître les principes de base de consultation du World Wide Web ainsi que les spécificités du navigateur qu'il utilise. Il est donc inutile de chercher dans la documentation ci-jointe les informations relevant de ces connaissances.

2. Notion de session

Au moment précis où vous avez activé FRANTEXT en cliquant sur le bouton "ACTIVATION DE FRANTEXT" de la page d'accueil, vous avez initialisé une session de travail.
Durant une session de travail, vous aurez la possibilité de créer des fichiers résidant sur l'espace disque du centre serveur. La session sera fermée et les fichiers seront automatiquement effacés au bout d'un délai de douze heures à compter du début de la session.

Si vous avez déjà débuté une session, et si vous revenez à la page d'accueil et cliquez à nouveau sur "ACTIVATION DE FRANTEXT", vous initialisez alors une nouvelle session. Vous êtes alors considéré comme un nouvel utilisateur, et les fichiers que vous avez créés dans la session précédente deviennent inaccessibles (à moins que vous ne reveniez dans une fenêtre correspondant à l'ancienne session en utilisant les possibilités éventuelles de retour en arrière offertes par votre logiciel de consultation).
Une telle acrobatie ne présente aucun intérêt. Il est donc déconseillé d'activer simultanément plusieurs sessions.

3. Principe général de FRANTEXT

FRANTEXT peut se définir comme un ensemble comportant :

- Un important corpus de textes français, du XVI ème au XX ème siècle, saisis sur support informatique. Le corpus est constitué d'environ 3500 oeuvres (soit plus d'un milliard de caractères). Il contient à peu près 80% d'oeuvres littéraires et 20% d'ouvrages techniques illustrant les diverses disciplines scientifiques.
- Un logiciel de consultation.

Le principe de base du logiciel est très simple. On peut distinguer deux phases fondamentales :

1. **Choix des textes que l'on veut étudier.**

Le choix des textes à étudier se fait grâce au service de sélection bibliographique. Il permet de sélectionner tout ou partie des textes disponibles dans
FRANTEXT grâce à des critères bibliographiques (titres, auteurs, dates, genres littéraires).
L'ensemble des textes ainsi sélectionnés est appelé corpus de travail.

2. **Série d'études portant sur les textes choisis.**

Une fois le corpus de travail établi, l'utilisateur dispose d'une large batterie d'outils lui permettant de procéder à diverses études. Le corpus de travail est mémorisé à l'issue de l'opération de sélection bibliographique. Toute étude porte donc implicitement sur le plus récent corpus de travail que l'on a défini.
Les services sont accessibles via un système de menus qui part d'un "menu principal".

Chaque entrée d'un menu est constituée :
D'une touche : le fait de cliquer sur cette touche fournit une description du service.
D'un texte correspondant au nom du service : le fait de cliquer sur ce texte déclenche l'activation du service. Ceci a pour effet de faire apparaître un formulaire sur l'écran. A chaque type de service correspond un type de formulaire.

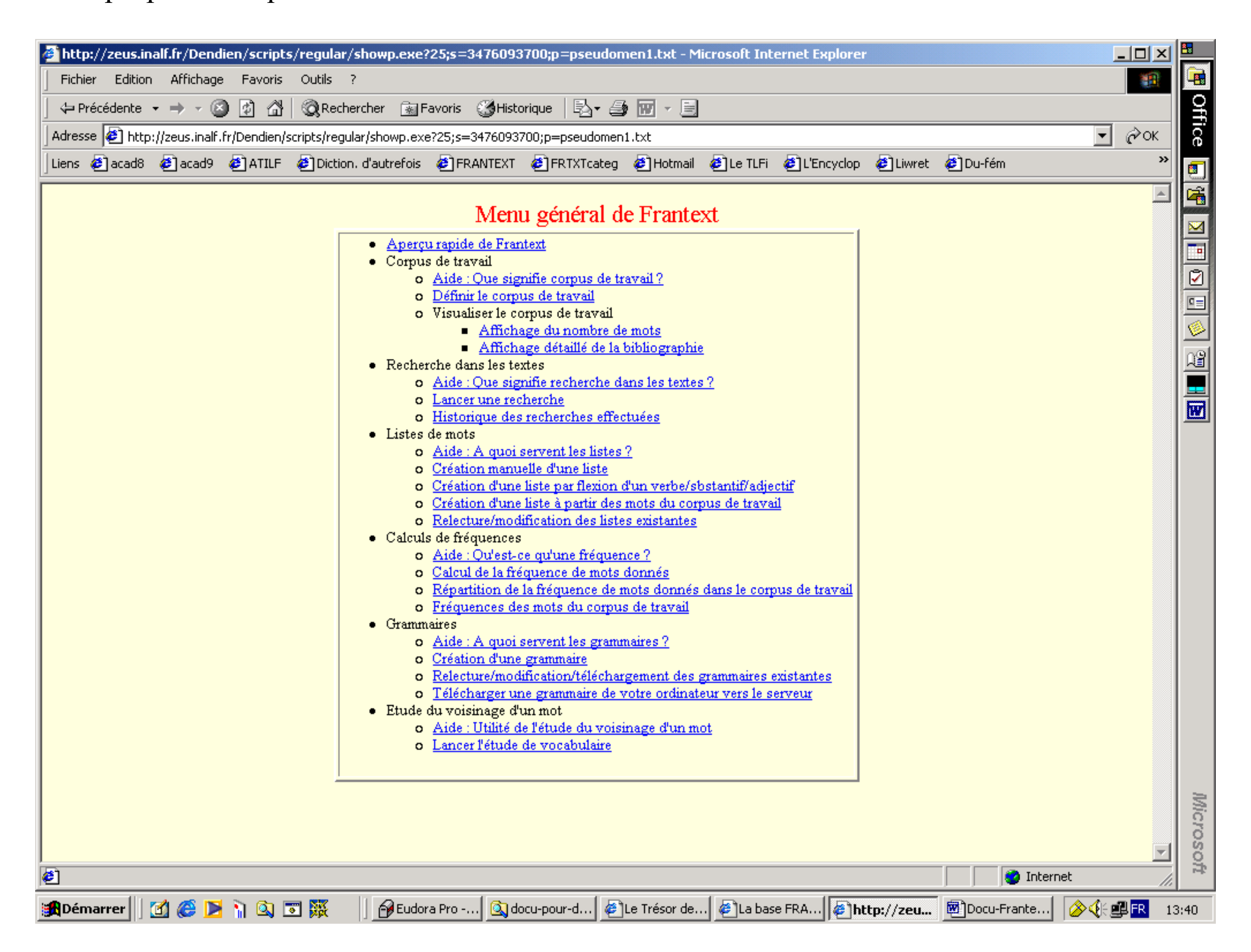
L'utilisateur doit remplir le formulaire en fonction de ses désirs, en se laissant guider par les explications qu'il contient ou par les touches d'aide qui s'y trouvent. Après remplissage du formulaire, il suffit de cliquer sur un bouton déclenchant l'exécution du service correspondant (ce bouton est très clairement mis en évidence dans les formulaires).
Dès la fin de l'exécution du service, l'utilisateur recevra des pages de résultats qu'il pourra à loisir consulter ou sauvegarder avant de revenir aux menus pour accéder à d'autres services.

Les différents services de FRANTEXT sont assez bien auto-documentés pour qu'un plus long discours soit ici nécessaire!
A vous de jouer en allant au menu principal.

Note : : Frantext non catégorisé contient 3665 textes
Frantext catégorisé contient 1940 textes.

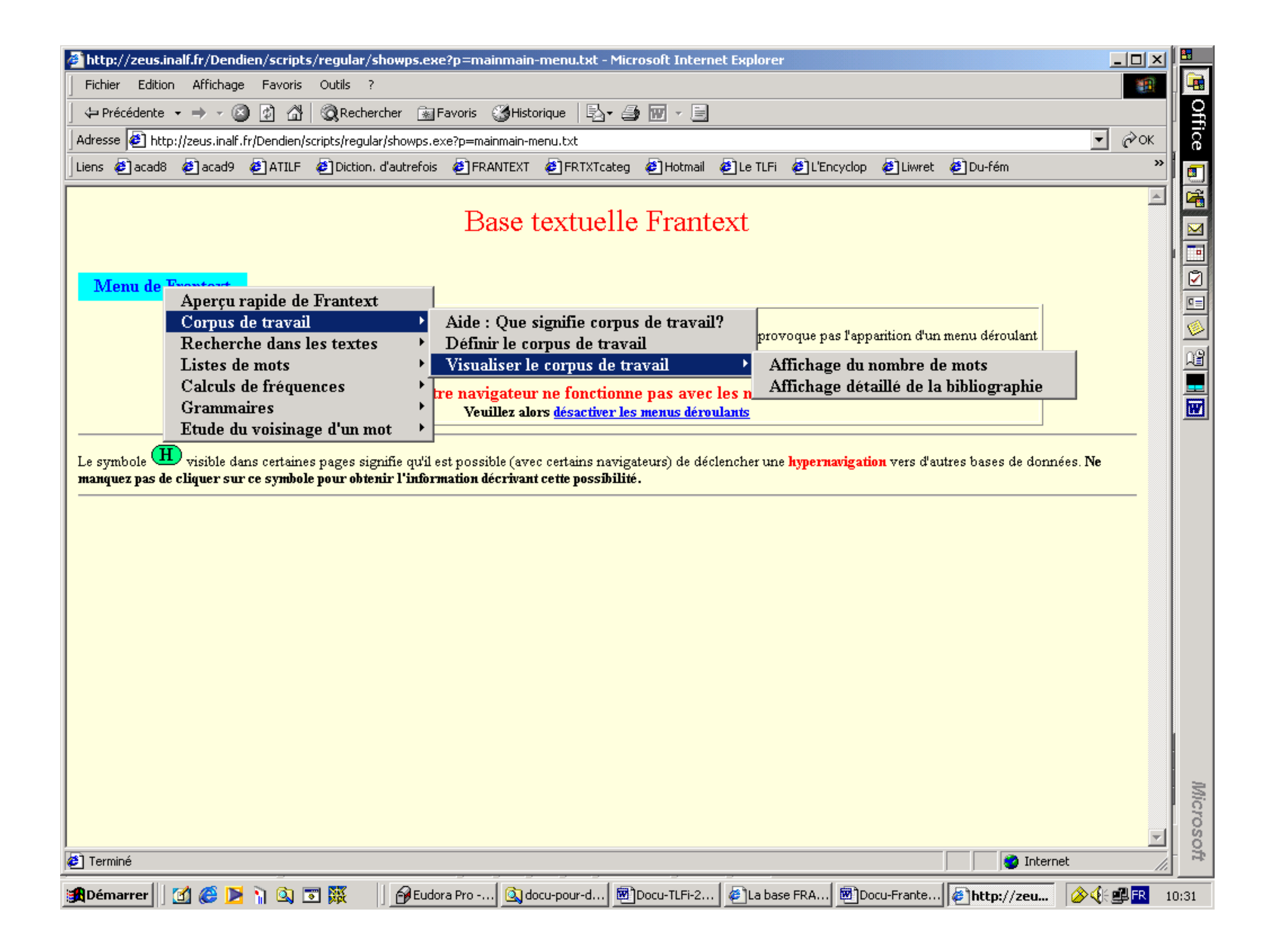
4. Menu principal de FRANTEXT

Ce menu est celui proposé dans le formulaire de requête (Recherche dans les textes).
Il est proposé lorsqu'on désactive les menus déroulants.



Ajoutons que l'on peut trouver des indications sur Visualisation et Rapatriement des résultats, par des boutons figurant dans les résultats des requêtes. Voir table des matières.

Corpus de travail



1. Aide : Que signifie corpus de travail

La base Frantext vous offre un certain nombre de services opérant sur tout ou partie des textes disponibles. Citons par exemple :

- Recherche dans les textes (d'un mot, d'une expression)
- Calcul de fréquences (comptage du nombre d'occurrences d'un mot ou de plusieurs mots donnés)
- Etc.

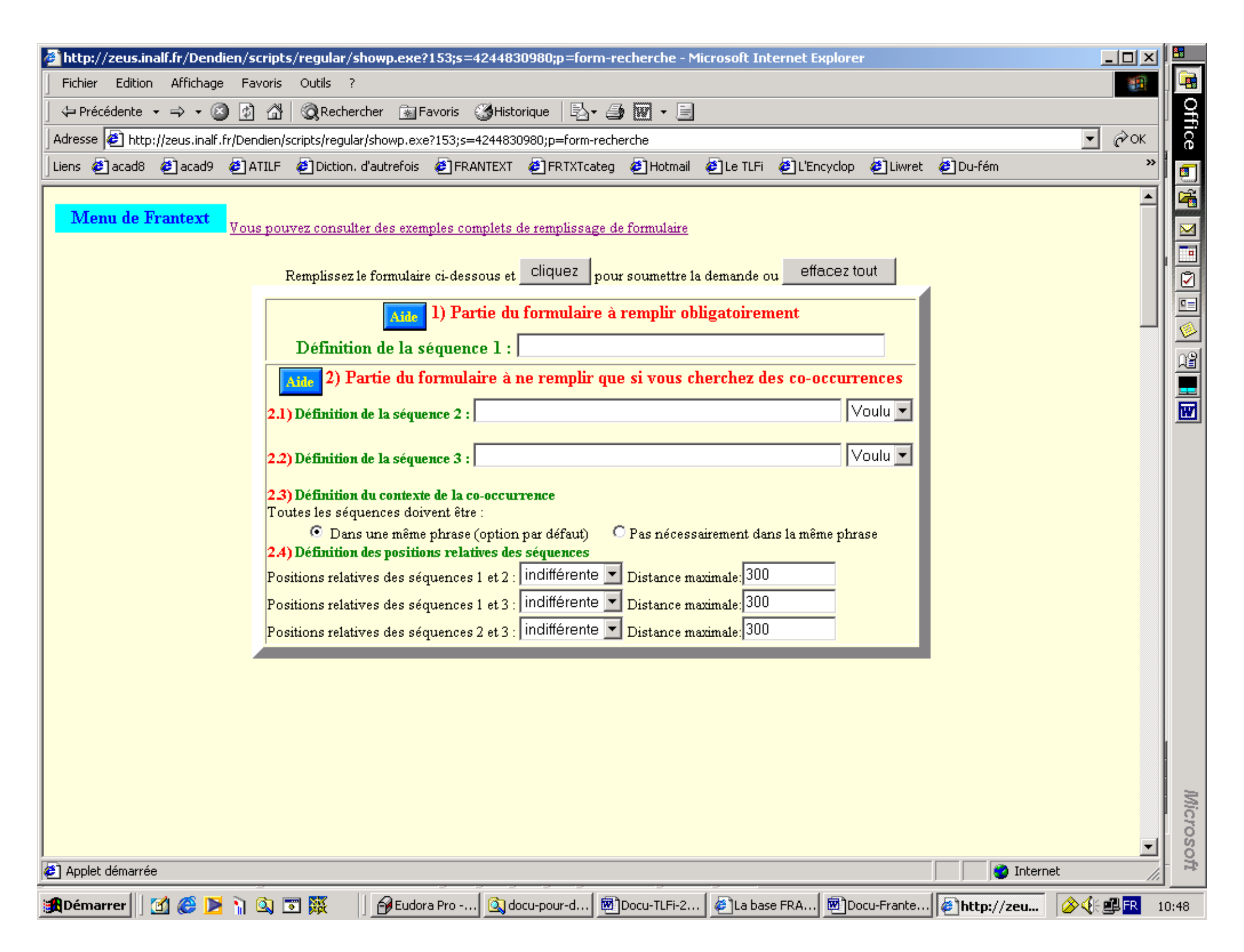
Définir votre corpus de travail, c'est choisir sur quels textes vous voulez travailler.

Le choix du corpus de travail peut se faire selon les critères suivants : Auteurs, Dates d'écriture, Œuvres littéraires.

Lorsque vous définissez le corpus de travail, votre choix est enregistré : dès lors, tous les services de Frantext opéreront sur ce choix. Vous pouvez cependant à tout moment, et aussi souvent que vous le désirez, définir un nouveau corpus de travail : il se substituera à l'ancien.

Si vous désirez un affichage de la bibliographie des textes du corpus de travail, vous pouvez demander une visualisation du corpus de travail.

2. Définir le corpus de travail



- ## Saisie du nom des auteurs

Il est possible d'entrer un seul ou plusieurs noms. Si plusieurs noms sont donnés, ils doivent être séparés par des virgules.
Exemple : Vous entrez *zola,flaubert,hugo*
Les noms peuvent être tapés en minuscules, sans caractères accentués.
Exemple : Les noms peuvent être donnés de manière fragmentaire, mais danger : *sand* appellera Sand et Sandeau. Pour avoir seulement Sand, taper *sand:george* (en un mot)
Exemple : *auber* suffit pour sélectionner Flaubert.

- ## Saisie du titre

Entrer un titre consiste en fait à entrer une chaîne de caractères.
Tous les ouvrages dont le titre contient cette chaîne seront sélectionnés.
Exemple : Vous entrez *dame*
Ceci sélectionne *"Au bonheur des dames", "Notre Dame de Paris"*, etc...
La chaîne peut être tapée en minuscules, sans caractères accentués

- ### Saisie du genre

Si aucun genre n'est coché, le critère genre est considéré comme inutilisé : tous les ouvrages seront sélectionnés quel que soit leur genre.
Si au moins un genre est coché, seuls les ouvrages correspondant au(x) genre(s) coché(s) seront sélectionnés.
Le résultat de la sélection précédente, sur « Leroux » a été :

3. Visualiser le corpus de travail

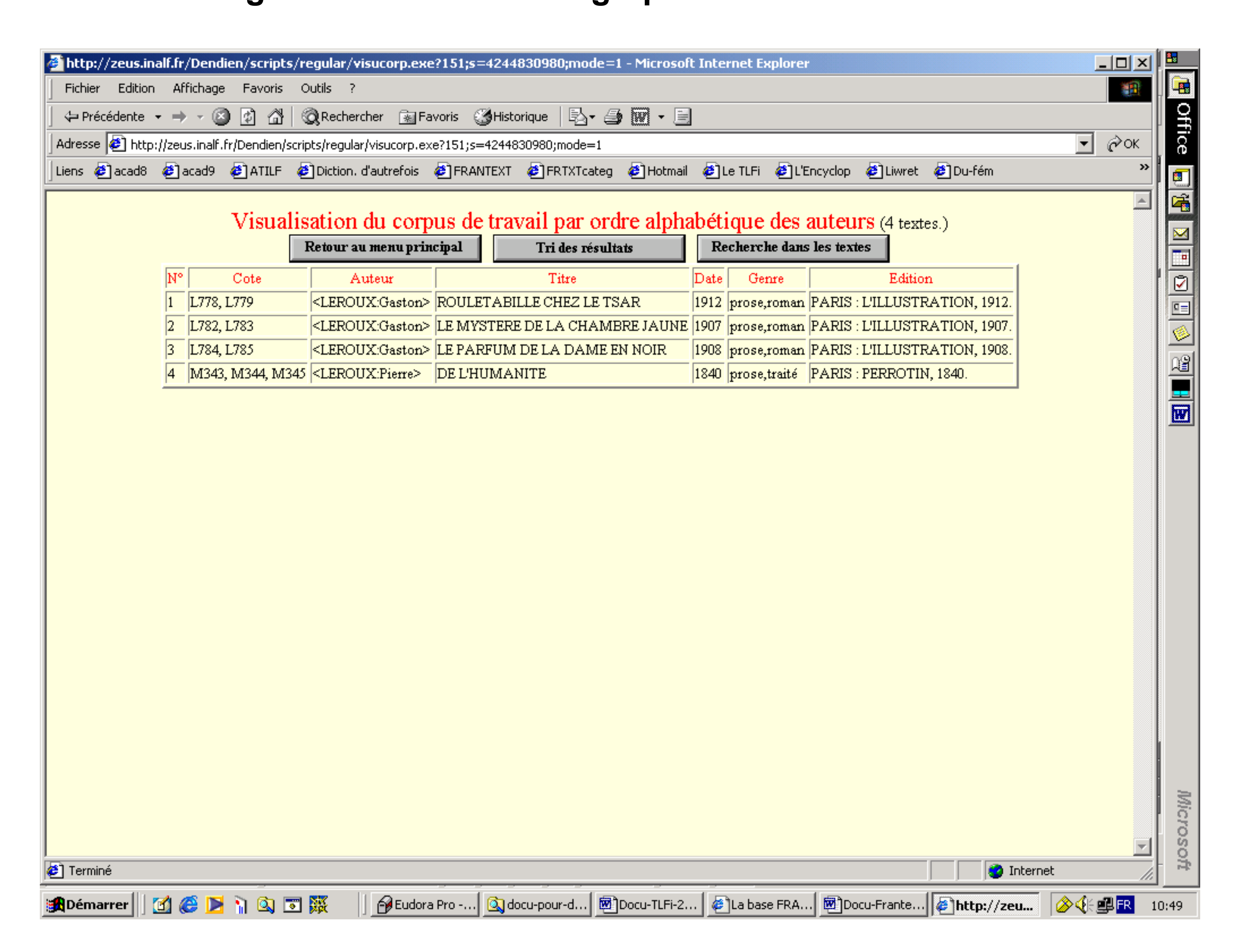
- ### Réponse à la requête

Nombre de textes sélectionnés : 4

- ### Affichage du nombre de mots

Nombre d'occurrences dans le corpus : 593032

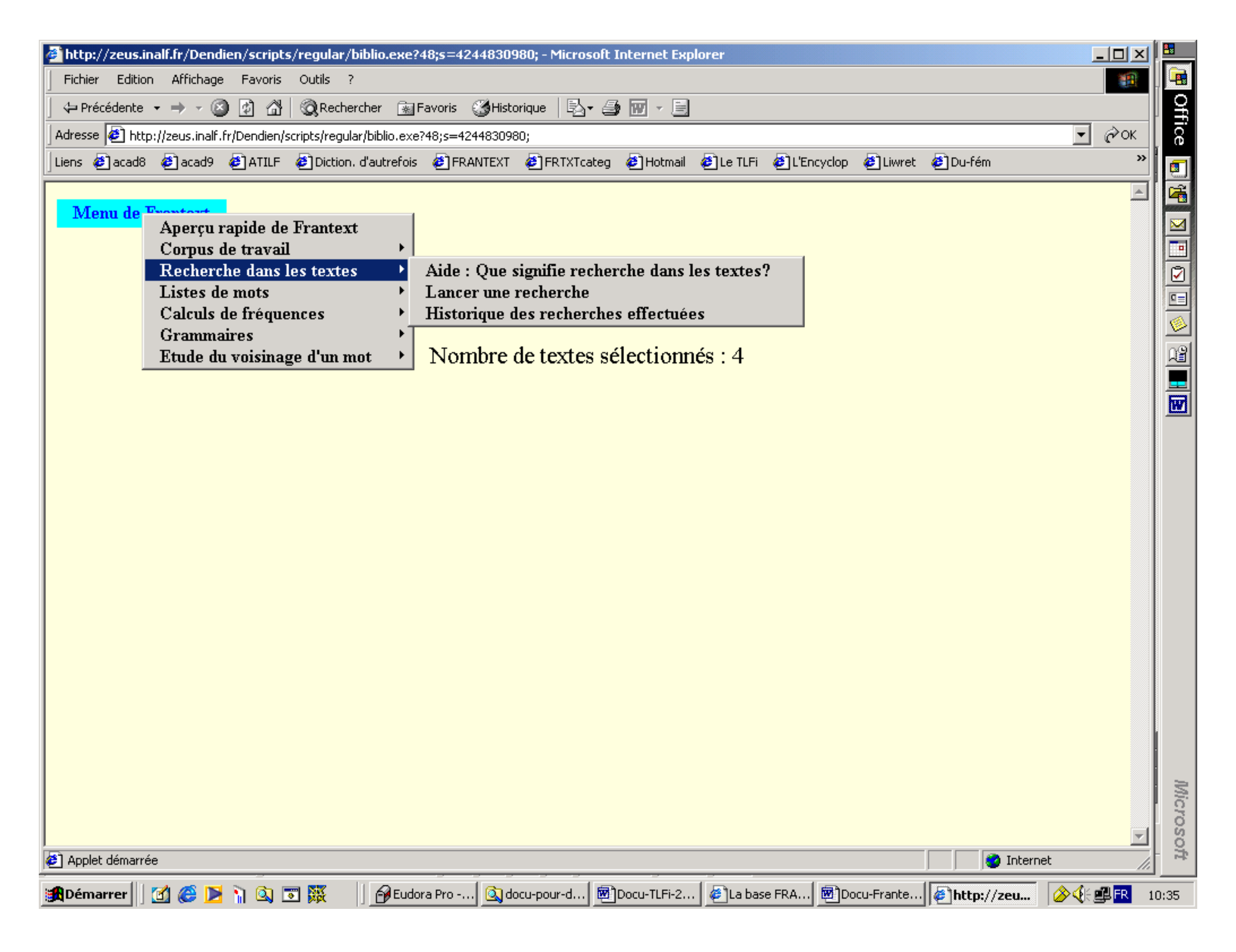
- ### Affichage détaillé de la bibliographie



N°	Cote	Auteur	Titre	Date	Genre	Edition
1	L778, L779	LEROUX:Gaston	ROULETABILLE CHEZ LE TSAR	1912	prose,roman	PARIS : L'ILLUSTRATION, 1912.
2	L782, L783	LEROUX:Gaston	LE MYSTERE DE LA CHAMBRE JAUNE	1907	prose,roman	PARIS : L'ILLUSTRATION, 1907.
3	L784, L785	LEROUX:Gaston	LE PARFUM DE LA DAME EN NOIR	1908	prose,roman	PARIS : L'ILLUSTRATION, 1908.
4	M343, M344, M345	LEROUX:Pierre	DE L'HUMANITE	1840	prose,traité	PARIS : PERROTIN, 1840.

Recherche dans les textes

Ce service consiste à rechercher **dans les textes du corpus de travail** des mots ou des expressions. L'objet de la recherche est défini en remplissant un formulaire. A l'issue d'une recherche, il est possible de visualiser les résultats (les mots ou expressions cherchés sont restitués dans leur contexte).
NOTA : Le nombre de résultats restitués est limité à **50000**.

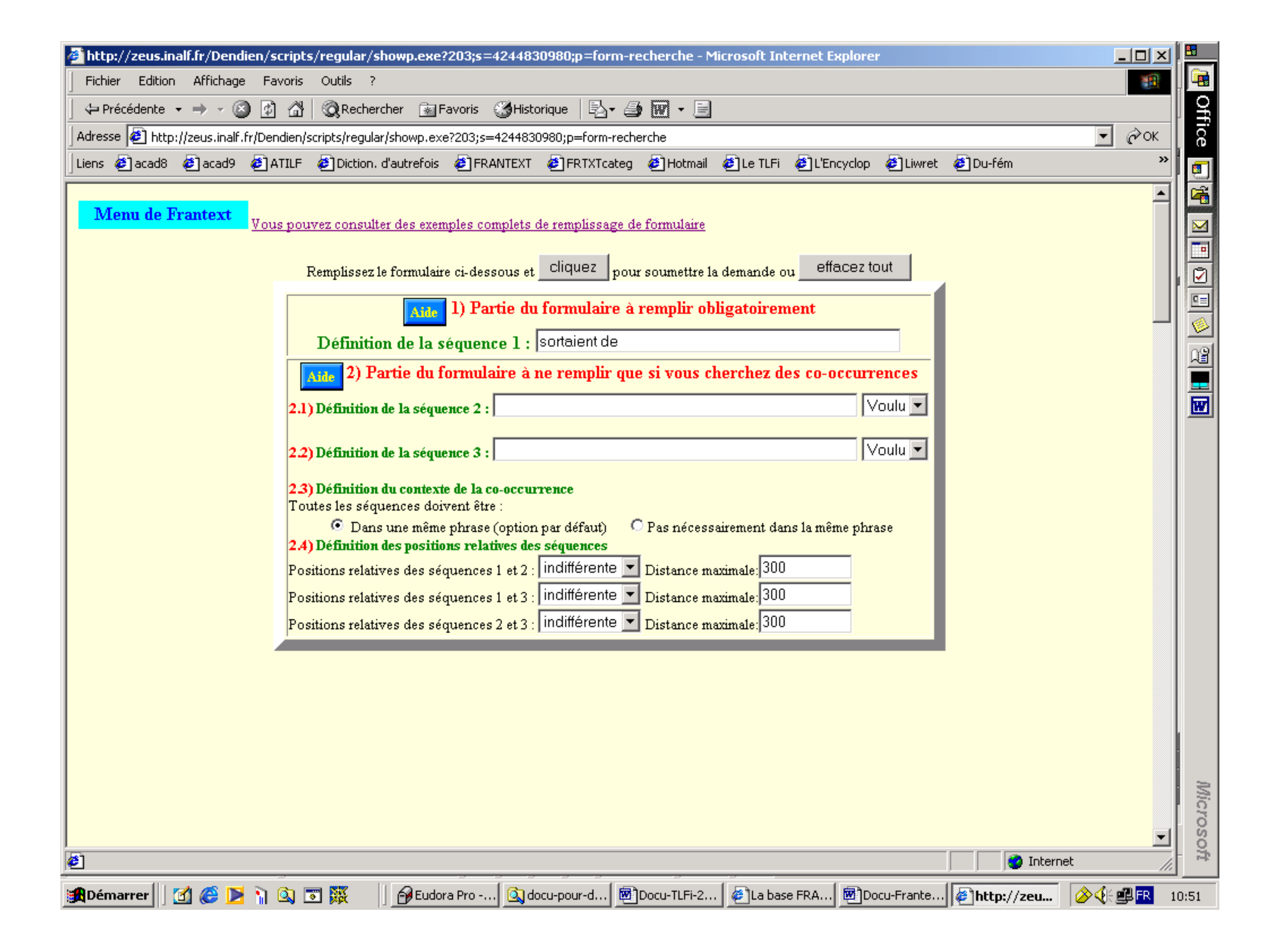


1. Que signifie Recherche dans les textes ?

Le terme **recherche dans les textes** désigne l'action qui consiste à rechercher dans le corpus de travail les contextes contenant un mot ou une séquence de mots donnés. Par exemple, on pourra rechercher des occurrences du mot *maison* ou de la séquence *maison blanche*. Il est également possible de rechercher des co-occurrences (apparition simultanée) de mots ou de séquences de mots, soit dans la même phrase, soit à une distance maximale donnée les uns des autres.

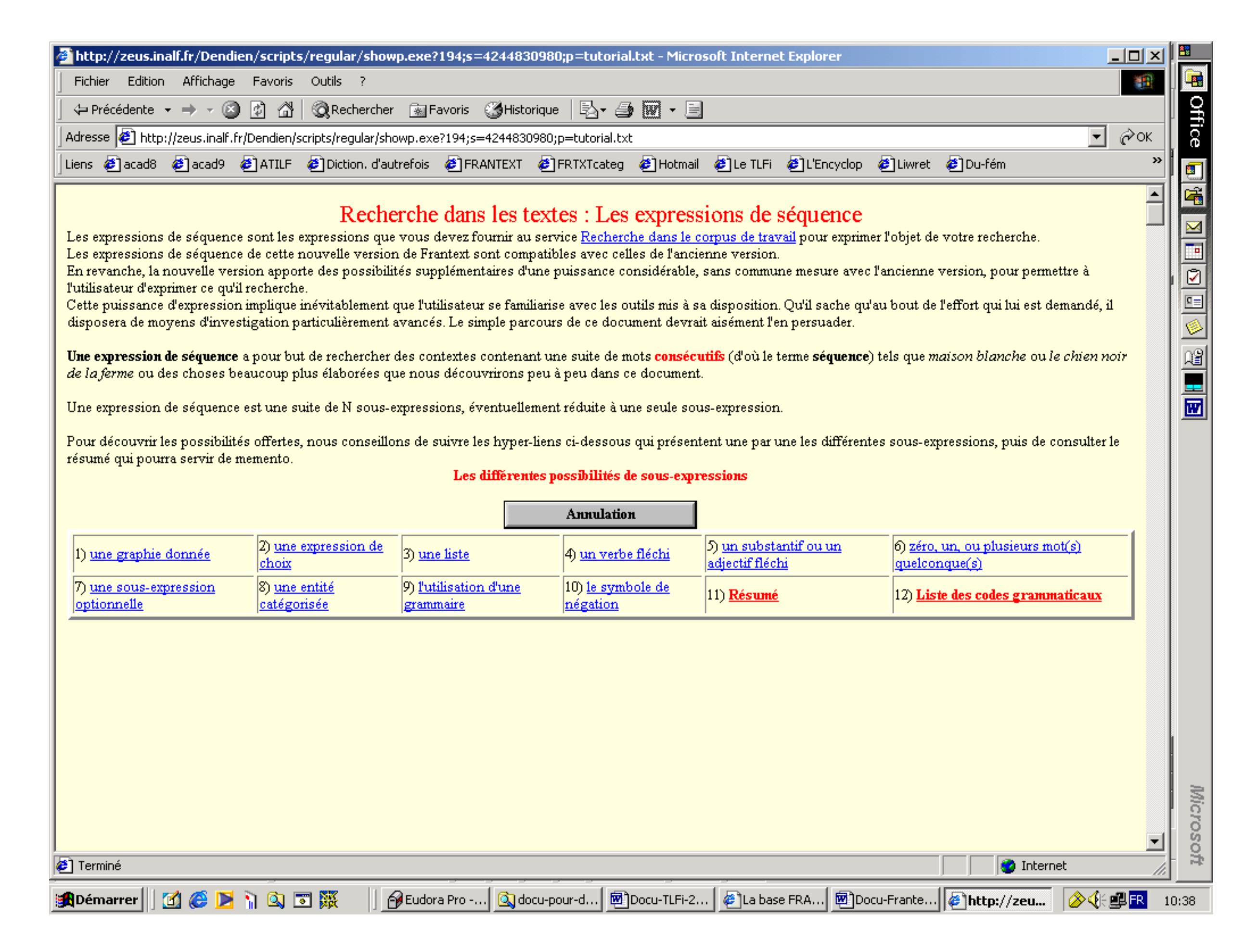
Ce qui différencie Frantext des autres bases textuelles existantes est que les possibilités de recherche ne s'arrêtent pas là. Frantext permet, grâce aux listes de mots, à la flexion automatique des verbes, substantifs ou adjectifs, et grâce aux grammaires paramétrables, d'exprimer des demandes d'une complexité arbitrairement élevée.

2. Lancer une recherche



Aide 1) : Partie du formulaire à remplir obligatoirement

Ceci renvoie aux expressions de séquence, voir ci-dessous §3



Les expressions de séquences font l'objet d'un chapitre particulier.
Les **points 8 et 9 et 12** (entités catégorisées, grammaires, et codes grammaticaux) font référence à la base Frantext catégorisée dont il sera question dans la deuxième section. Nous n'en parlerons donc pas ici.

Aide 2) : Partie du formulaire à remplir pour co-occurrences :

Dans le cas où vous remplissez non seulement le cadre *Séquence 1*, mais aussi le cadre *Séquence 2*, et éventuellement *Séquence 3*, c'est que vous désirez faire une co-occurrence de séquences.
Plus précisément, vous pouvez exprimer que *Séquence 2* ou *Séquence3* sont **voulues** dans le contexte de *Séquence 1* ou **exclues** de ce contexte.
Vous avez à préciser dans quels contextes les séquences 2 et 3 doivent être trouvées (option **voulu**) ou ne pas être trouvées (option **exclu**) en même temps que la séquence 1.

- Vous pouvez exiger que le contexte soit limité ou non à une même phrase.
- Vous pouvez exiger également que les séquences occupent un ordre imposé (séquence X avant ou après séquence Y) et soient à une distance maximale l'une de l'autre.

La distance maximale admissible est de **300 mots**.
La **distance entre deux séquences** est définie de la manière suivante :
Si on suppose tous les mots du texte numérotés 1, 2, 3,… la distance entre deux séquences est égale à la valeur absolue de la différence entre les numéros des premiers mots de chaque séquence.

Par exemple, dans le texte *bien faire et laisser dire*, la distance entre *bien faire* et *laisser dire* est égale à 3.

Exemples de remplissage de formulaire

Exemple 1 On désire rechercher le mot **cheval** au singulier.
Il suffit de taper **cheval** dans le cadre séquence 1 et de soumettre la demande.

Exemple 2
On désire rechercher le mot **cheval** au singulier ou au pluriel.
Il suffit de taper **&mcheval** dans le cadre séquence 1 et de soumettre la demande.

Exemple 3 On désire rechercher le mot **cheval** au singulier ou au pluriel lorsque ce mot apparaît dans la même phrase que le mot **selle**
Il suffit

- de taper **&mcheval** dans le cadre séquence 1
- de taper **selle** dans le cadre séquence 2

Soumettre ensuite la demande.

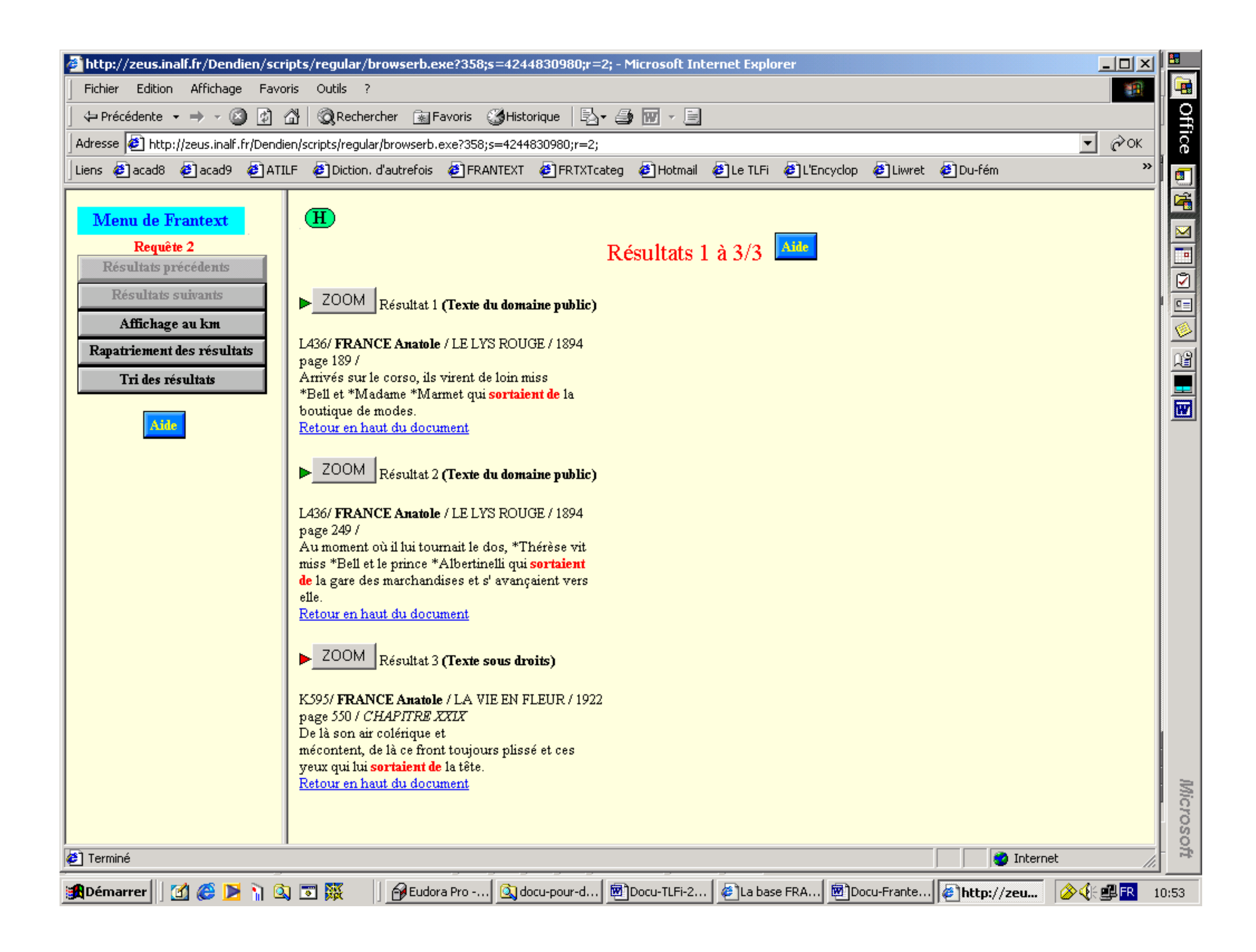
Exemple 4
On désire chercher les phrases qui contiennent le mot **femme**, sans que le mot **homme** n'apparaisse à droite du mot femme à moins de 10 mots. Il suffit

- de taper **femme** dans le cadre séquence 1
- de taper **homme** dans le cadre séquence 2 et de choisir sur la même ligne l'option **Exclu**
- de choisir dans "Positions relatives des séquences 1 et 2 " l'option **1 AVANT 2** et sur la même ligne de remplir le cadre "distance maximale" avec la valeur **10**

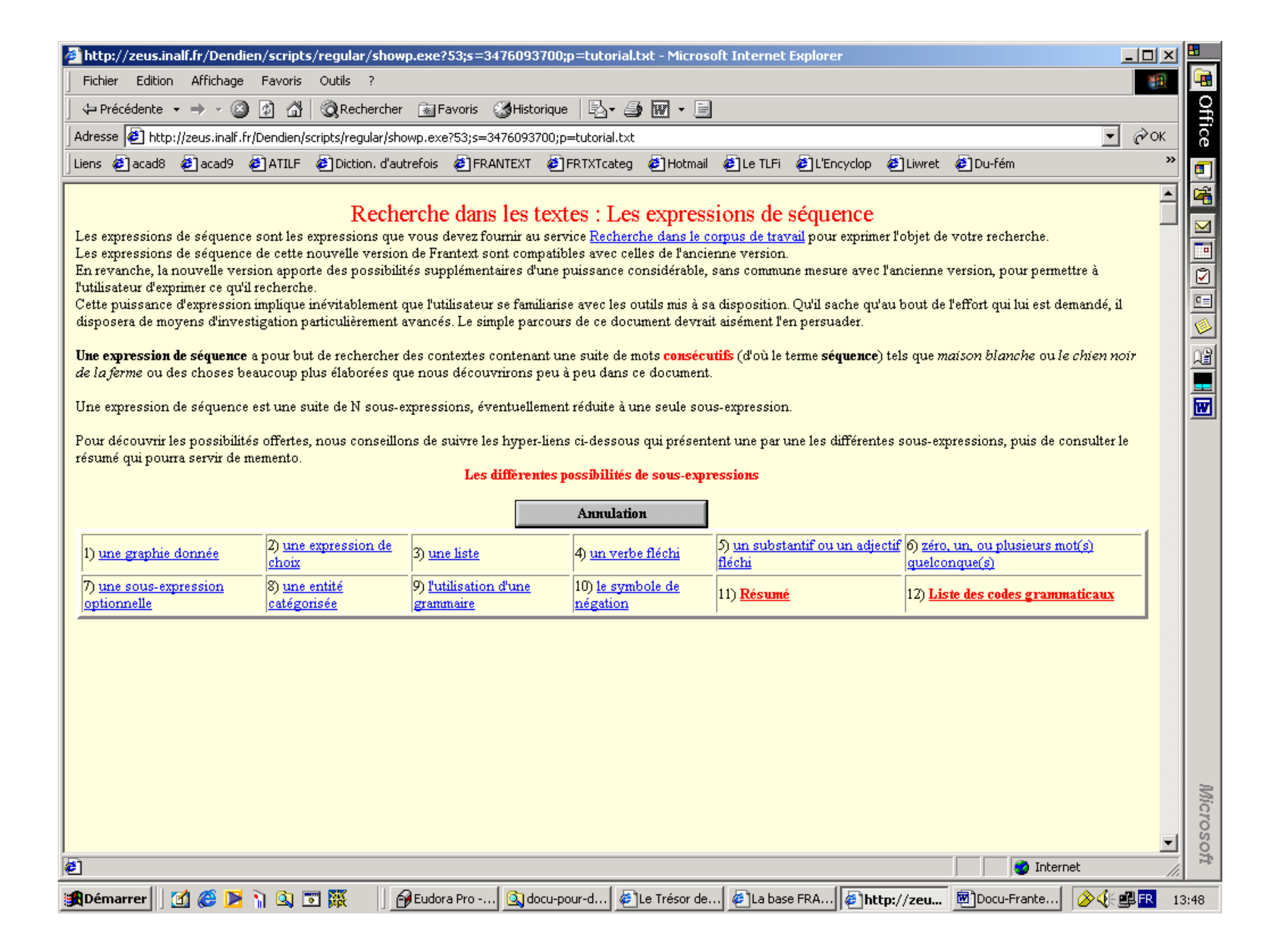
Nota : Pour une description détaillée de la manière de spécifier une séquence, on se reportera à la section Définition d'une séquence.

3. Exemple de résultat :

Corpus sélectionné : auteur = France -> 8 textes
Séquence demandée : *sortaient de* -> 3 résultats


http://zeus.inalf.fr/Dendien/scripts/regular/browserb.exe?358;s=4244830980;r=2; - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente Rechercher Favoris Historique
Adresse http://zeus.inalf.fr/Dendien/scripts/regular/browserb.exe?358;s=4244830980;r=2;
Liens acad8 acad9 ATILF Diction. d'autrefois FRANTEXT FRTXTcateg Hotmail Le TLFi L'Encyclop Livret Du-fém
Menu de Frantext
Requête 2
Résultats précédents
Résultats suivants
Affichage au km
Rapatriement des résultats
Tri des résultats
Aide
H
Résultats 1 à 3/3
Aide
ZOOM Résultat 1 (Texte du domaine public)
L436/ FRANCE Anatole / LE LYS ROUGE / 1894
page 189 /
Arrivés sur le corso, ils virent de loin miss
*Bell et *Madame *Marmet qui sortaient de la
boutique de modes.
Retour en haut du document
ZOOM Résultat 2 (Texte du domaine public)
L436/ FRANCE Anatole / LE LYS ROUGE / 1894
page 249 /
Au moment où il lui tournait le dos, *Thérèse vit
miss *Bell et le prince *Albertinelli qui sortaient
de la gare des marchandises et s' avançaient vers
elle.
Retour en haut du document
ZOOM Résultat 3 (Texte sous droits)
K595/ FRANCE Anatole / LA VIE EN FLEUR / 1922
page 550 / CHAPITRE XXIX
De là son air colérique et
mécontent, de là ce front toujours plissé et ces
yeux qui lui sortaient de la tête.
Retour en haut du document
Terminé
Internet
Démarrer
Eudora Pro -...
docu-pour-d...
Docu-TLFi-2...
La base FRA...
Docu-Frante...
http://zeu...
10:53
Office
Microsoft

Les expressions de séquences



Recherche dans les textes : Les expressions de séquence

Les expressions de séquence sont les expressions que vous devez fournir au service Recherche dans le corpus de travail pour exprimer l'objet de votre recherche.
Les expressions de séquence de cette nouvelle version de Frantext sont compatibles avec celles de l'ancienne version.
En revanche, la nouvelle version apporte des possibilités supplémentaires d'une puissance considérable, sans commune mesure avec l'ancienne version, pour permettre à l'utilisateur d'exprimer ce qu'il recherche.
Cette puissance d'expression implique inévitablement que l'utilisateur se familiarise avec les outils mis à sa disposition. Qu'il sache qu'au bout de l'effort qui lui est demandé, il disposera de moyens d'investigation particulièrement avancés. Le simple parcours de ce document devrait aisément l'en persuader.

Une expression de séquence a pour but de rechercher des contextes contenant une suite de mots **consécutifs** (d'où le terme **séquence**) tels que *maison blanche* ou *le chien noir de la ferme* ou des choses beaucoup plus élaborées que nous découvrirons peu à peu dans ce document.

Une expression de séquence est une suite de N sous-expressions, éventuellement réduite à une seule sous-expression.

Pour découvrir les possibilités offertes, nous conseillons de suivre les hyper-liens ci-dessous qui présentent une par une les différentes sous-expressions, puis de consulter le résumé qui pourra servir de memento.

Les différentes possibilités de sous-expressions

Annulation

1) une graphie donnée	2) une expression de choix	3) une liste	4) un verbe fléchi	5) un substantif ou un adjectif fléchi	6) zéro, un, ou plusieurs mot(s) quelconque(s)
7) une sous-expression optionnelle	8) une entité catégorisée	9) l'utilisation d'une grammaire	10) le symbole de négation	11) **Résumé**	12) **Liste des codes grammaticaux**

1. Une graphie donnée

C'est la manière la plus "banale" de désigner un élément de séquence.

Exemple d'expression de séquence composée de deux graphies données:
maison blanche
Cette expression recherche des occurrences de ***maison blanche***.

2. Une expression de choix

Une expression de choix s'exprime sous la forme : (Choix$_1$ | Choix$_2$ | ... | Choix$_n$)
dans laquelle Choix$_1$, Choix$_1$, ..., Choix$_n$ sont des expressions de séquence.

Exemple 1 d'utilisation d'expression de choix :
maison (blanche | bleue)

Remarques :

- L'expression ci-dessus est équivalente à **(maison blanche | maison bleue)**. On notera cependant que la mise en facteur du mot "maison" a le double avantage de donner une expression plus compacte et de procurer un temps de recherche plus bref lors de l'exécution de la requête.
- Des caractères blancs ont été ajoutés dans les exemples ci-dessus pour les rendre plus lisibles. On aurait pu les supprimer et écrire, par exemple, **maison(blanche|bleue)**

Exemple 2 :
Puisqu'une expression de choix est une expression de séquence et qu'un des choix est lui-même une expression de séquence, on en déduit que l'on peut imbriquer les expressions de choix à volonté.
(maison|palais)d'un(blanc(immaculé|sale)|bleu(d'azur|profond))
est une expression qui cherchera les occurrences de ***maison d'un blanc immaculé***, ***palais d'un blanc immaculé***, ***maison d'un blanc sale***, ***palais d'un blanc sale***, ***maison d'un bleu d'azur***, ***palais d'un bleu d'azur***, ***maison d'un bleu profond***, ***palais d'un bleu profond***

3. Les listes

&lxxx désigne un des mots de la liste xxx.
(On se reportera au service création/édition de listes de mots pour obtenir une description des différentes manières de créer une liste de mots.)
Exemple :
Si dans une liste de nom *couleur* on a les mots rouge, vert, jaune, on peut écrire l'expression de séquence suivante :
volet &lcouleur
pour chercher *volet rouge, volet vert, volet jaune*

Remarque : une liste est une suite d'éléments constitués d'une seule graphie (le « blanc » est considéré comme un séparateur de mots). Un élément de liste ne peut donc pas être par exemple « vert bouteille ». Cette restriction est compensée par le recours aux grammaires présentées plus loin dans ce document.

4. Verbe fléchi

&cxxx
désigne une des formes fléchies du verbe dont l'infinitif est xxx
Exemple d'expression de séquence :
&caimer les bons repas

5. Substantif ou adjectif fléchi

&mxxx
désigne une des formes d'un substantif dont le singulier est xxx, d'un adjectif dont le masculin singulier est xxx.
Exemple d'expression de séquence :
&mfleur &mvert
Cette séquence trouvera les occurrences de ***fleur verte*** ou ***fleurs vertes***

6. Zéro, un ou plusieurs mot(s) quelconque(s)

&q(n_1,n_2)
désigne une suite de mots dont le nombre est compris entre n_1 et n_2 .
Exemple d'expression de séquence :
un &q(0,2) homme
Cette expression de séquence pourra trouver des contextes tels que ***un*** soit séparé de ***homme*** par zéro, un ou deux mots, donc des contextes tels que ***un homme***, ***un grand homme***, ***un très petit homme***.
Notes :

- Les nombres n_1 et n_2 doivent respecter les règles suivantes :
  - n_1 doit être positif ou nul. n_2 doit être strictement positif.
  - n_2 doit être supérieur à n_1
  - La différence n_2-n_1 doit être inférieure ou égale à 7 (ce qui permet une amplitude de 8 mots). Cette restriction est due au souci de maintenir un temps de recherche raisonnable en évitant au logiciel d'avoir à faire des hypothèses trop nombreuses. Il est demandé de ne pas essayer de contourner cette restriction avec une expression du genre **un&q(0,7) &q(0,7)homme** qui conduirait effectivement à tolérer jusqu'à 16 mots entre ***un*** et ***homme***, mais au prix d'une violente dégradation du temps de réponse dont l'utilisateur serait la première victime. La bonne solution pour chercher des séquences distantes de plus de 8 mots est de chercher deux séquences en cooccurrence (voir aide sur cette rubrique, cadre 2 du formulaire de recherche).
- L'expression **&q** utilisable dans les versions antérieures et qui désigne un mot quelconque est toujours utilisable. Elle est strictement équivalente à **&q(1,1)**.

7. Sous expressions optionnelles

Supposons que l'on veuille chercher des contextes contenant ***un homme*** ou ***un grand homme***.
Une telle recherche peut s'exprimer avec **(un | un grand) homme** ou avec **un (homme | grand homme)** en utilisant une expression de choix.
De telles expressions sont lourdes et inélégantes.

Nous proposons donc de les simplifier avec le symbole **&?** dont la signification est la suivante :
Si **&?** est situé devant une expression de séquence, alors cette expression est **facultative** dans les contextes recherchés.

Exemples :

- **un &?grand homme** recherchera les contextes ***un homme*** ou ***un grand homme***.
- **un &?(très grand) homme** recherchera les contextes ***un homme*** ou ***un très grand homme***.
- **un &?(&?très grand) homme** recherchera les contextes ***un homme*** ou ***un grand homme*** ou ***un très grand homme***.

8. Les entités catégorisées

renvoient à la base textuelle catégorisée

9. Les grammaires

Préambule

Les grammaires, telles que nous allons les définir, permettent de formuler de puissantes expressions de recherche capables de localiser dans un corpus des occurrences de phénomènes multiformes : par exemple, une référence à une période peut prendre la forme ***dès 1954***, ou ***en juillet dernier/prochain***, ***la semaine dernière/prochaine***, etc.
Les grammaires (au sens où nous l'entendons, qui est celui de la théorie des langages) ne sont pas à confondre avec les grammaires des langues naturelles. En revanche, elles permettent de rechercher des phénomènes (par exemple une construction pronominale de verbe) qui peuvent se manifester sous des formes très variables en raison des règles syntaxiques du français. En ce sens, il y a bien un lien conceptuel entre les grammaires que nous proposons et la grammaire française.
Malgré ce lien, les linguistes et les grammairiens comprendront que :

- nous donnons aux termes **grammaire** et **règle de grammaire** une signification différente de celle à laquelle ils sont habitués, mais qui est conforme à la terminologie de la théorie des langages.
- nous ne prétendons pas résoudre les difficultés des langues naturelles en proposant un simple outil : ce n'est pas parce que nous offrons une paire de ciseaux (nos grammaires) que nous prétendons être tailleurs (grammairiens)

Qu'est-ce qu'une grammaire ?

Une grammaire est un fichier qui contient des **règles**. Ce fichier peut être saisi par l'utilisateur grâce au service de création de grammaire.

L'intérêt des grammaires est triple :

- Une grammaire est un recueil d'expression de séquences tapées une fois pour toutes par l'utilisateur.
- Les expressions de séquences peuvent être paramétrées, ce qui permet de les utiliser pour des recherches multiples.
- Une grammaire permet d'élaborer facilement des expressions de séquences d'une complexité arbitrairement élevée.

Chaque règle de grammaire a un **nom** et un **corps**.

Un premier exemple de grammaire (voir autres exemples dans la base catégorisée)

Voici un exemple de grammaire : (**En rouge : nom de la règle**, **en vert : corps de la règle**)

quantifieur :
  très | assez | excessivement
qualifieur :
  grand | gros | petit
qualification :

&rqualifieur | &rquantifieur &rqualifieur

Supposons que vous ayez saisi une telle grammaire dans un fichier de nom **xxx**. Alors, vous pourrez utiliser les règles de cette grammaire dans des expressions de recherche lorsque vous ferez appel au service recherche dans le corpus de travail. Pour utiliser une règle de la grammaire xxx (règle "quantifieur" par exemple), vous taperez **&rquantifieur,xxx**. Une telle expression est appelée **invocation** de la règle "quantifieur" de la grammaire xxx.
Elle est équivalente à l'expression obtenue en plaçant le corps de la règle entre parenthèses :

&rquantifieur,xxx<=>(très | assez | excessivement)

Vous pourrez par exemple, avec la séquence **&rquantifieur,xxx &rqualifieur,xxx homme** chercher des contextes contenant ***très grand homme***, ***assez petit homme***, etc...

De même, en utilisant la règle **qualification** vous pourrez, avec la séquence **&rqualification,xxx homme** chercher des contextes contenant ***très grand homme*** (ou simplement ***grand homme***), ***assez petit homme*** (ou simplement ***petit homme***), etc...
Il est à noter que la règle **qualification** invoque les règles **quantifieur** et **qualifieur** sans spécifier de nom de grammaire : en effet le nom de la grammaire est facultatif à l'intérieur d'une même grammaire.

Notons qu'il existe un chapitre consacré aux Grammaires applicables à la base non catégorisée. Voir Table des matières.

10. Le symbole de négation

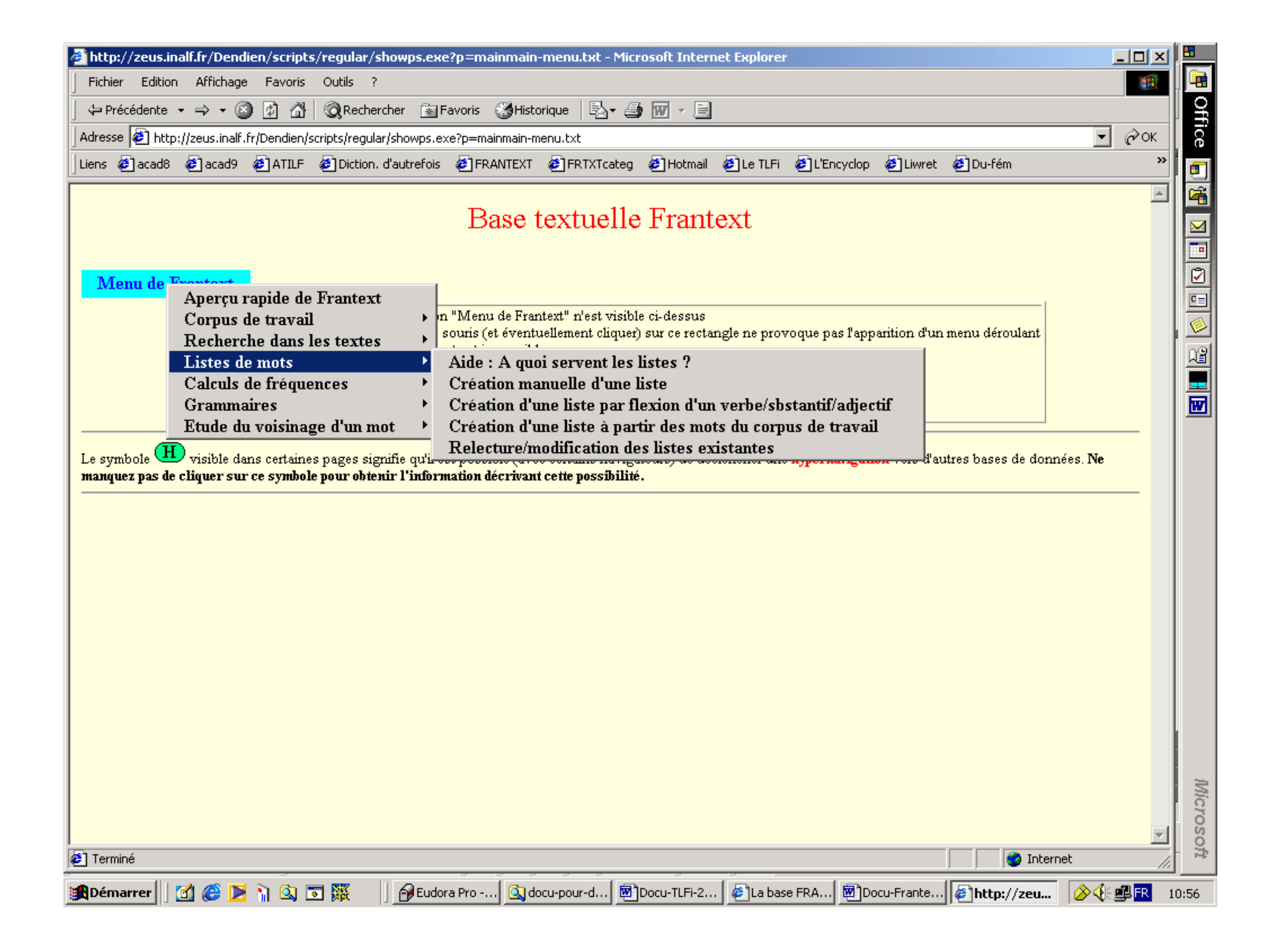
Le symbole de négation est le caractère ^. Il peut être placé devant n'importe quel type de sous expression. Quel que soit le type de sous-expression, l'ensemble formé par le symbole de négation et la sous-expression désigne **une graphie** qui est tout sauf le point de départ de la sous-expression.

Exemples :

- **homme ^très grand qui** est une expression de séquence qui va chercher tous les contextes du genre ***homme XXX grand qui*** tels que XXX ne soit pas égal à ***très***.
- **homme ^(très grand)** est une expression de séquence qui va chercher tous les contextes du genre ***homme XXX*** tels que ou bien XXX est différent de ***très***, ou bien XXX est égal à ***très*** et le mot suivant est différent de ***grand*** (autrement dit XXX n'est pas le point de départ d'une séquence ***très grand***).

Rappel : Nous ne parlons pas ici des points intitulés « **Entités catégorisées** » ni ni « **Résumé** » ni « **Codes grammaticaux** » ni de certains aspects de l'utilisation des « **Grammaires** », qui concernent la base catégorisée.

Listes de mots



1. Aide : A quoi servent les listes de mots ?

Les listes de mots sont des fichiers que vous créez et qui sont stockés sur le centre serveur. Ils resteront disponibles tant que votre session FRANTEXT est en cours. Les listes de mots que vous créez peuvent être utilisées comme des éléments textuels dont vous cherchez des occurrences (Voir le service « Recherche dans les textes »).

Pour créer une liste de mots, vous avez différents moyens à votre disposition.

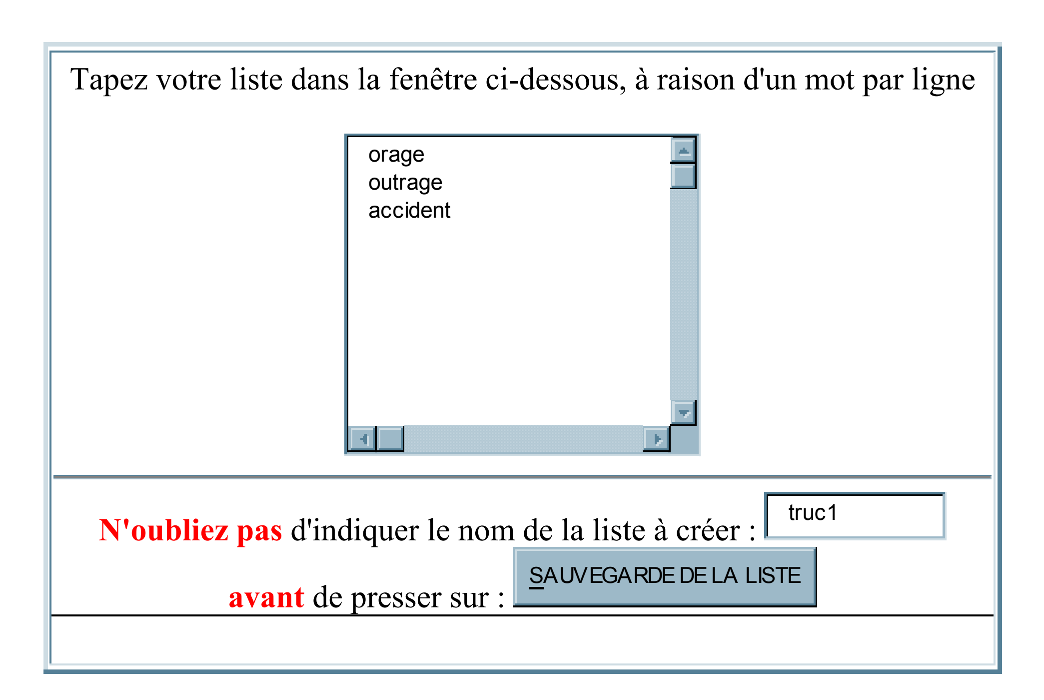
- Création manuelle : vous aurez à taper un par un les mots constituant la liste.
- Création automatique d'une liste par flexion d'un verbe (conjugaison), d'un substantif (forme singulier/pluriel) ou d'un adjectif (formes féminin/masculin, singulier/pluriel)
- Création automatique à partir des mots du corpus de travail qui correspondent à un filtre que vous indiquez (par exemple les mots qui commencent par *vol* , qui se terminent par *isme*, qui commencent par *c* suivi d'un caractère quelconque suivi de *p*, etc.).

Les listes peuvent être utilisées :

- Pour effectuer une recherche dans les textes : par exemple, si vous avez créé une liste de nom *sentiment* contenant les mots amour et haine, vous pourrez utiliser cette liste pour rechercher les contextes contenant un des mots de la liste, ou les contextes contenant un des mots de la liste précédé et/ou suivi d'autres mots donnés.
- Pour effectuer un calcul de fréquences de mots donnés : vous pourrez obtenir le nombre d'occurrences des mots d'une liste donnée dans le corpus de travail.
- Pour étudier la répartition de la fréquence de mots donnés dans le corpus de travail. La fréquence de chaque mot d'une liste donnée peut se calculer auteur par auteur ou œuvre par œuvre ou par tranches de temps du corpus de travail.
- Pour faire une étude de voisinage de mots donnés qui consiste à relever des mots du corpus de travail qui apparaissent dans le voisinage à proximité (par exemple même phrase) des occurrences des mots de votre liste.

On voit donc que les listes de mots sont un outil d'investigation de Frantext particulièrement important.

2. Création manuelle d'une liste



3. Création d'une liste par flexion d'un verbe/substantif/adjectif

Veuillez indiquer ci-dessous le lemme à fléchir, c'est-à-dire :

- l'infinitif s'il s'agit d'un verbe
- le singulier s'il s'agit d'un substantif
- le masculin singulier s'il s'agit d'un adjectif

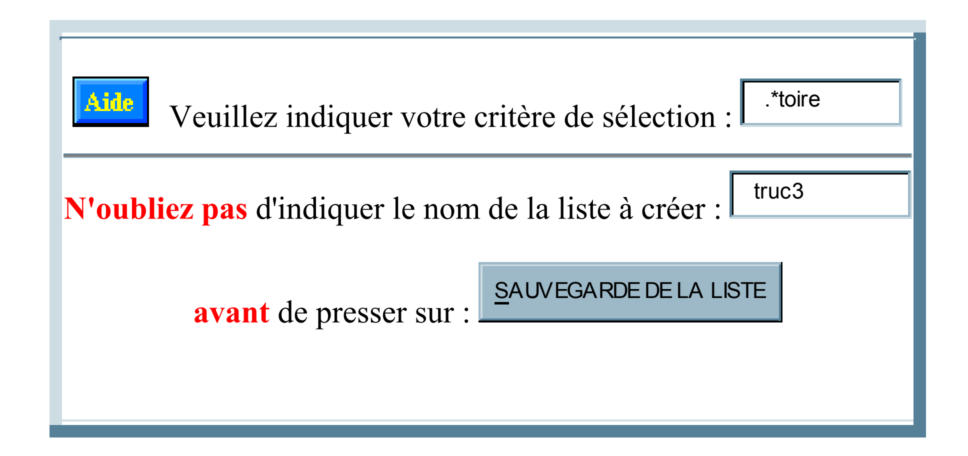
vert

N'oubliez pas d'indiquer le nom de la liste à créer : truc2

avant de presser sur : SAUVEGARDE DE LA LISTE

Lorsque je sauvegarde cette liste, un message m'avertit que la liste truc2 contient 4 graphies.

4. Création d'une liste à partir des mots du corpus



Lorsque je sauvegarde la liste, le message m'avertit que cette liste contient 14 graphies

les critères de sélection :

Le critère de sélection vous permet de définir quelles graphies du corpus de travail seront sélectionnées. Par exemple, le critère .*ismes? sélectionne tous les mots se terminant par *isme* ou *ismes*.

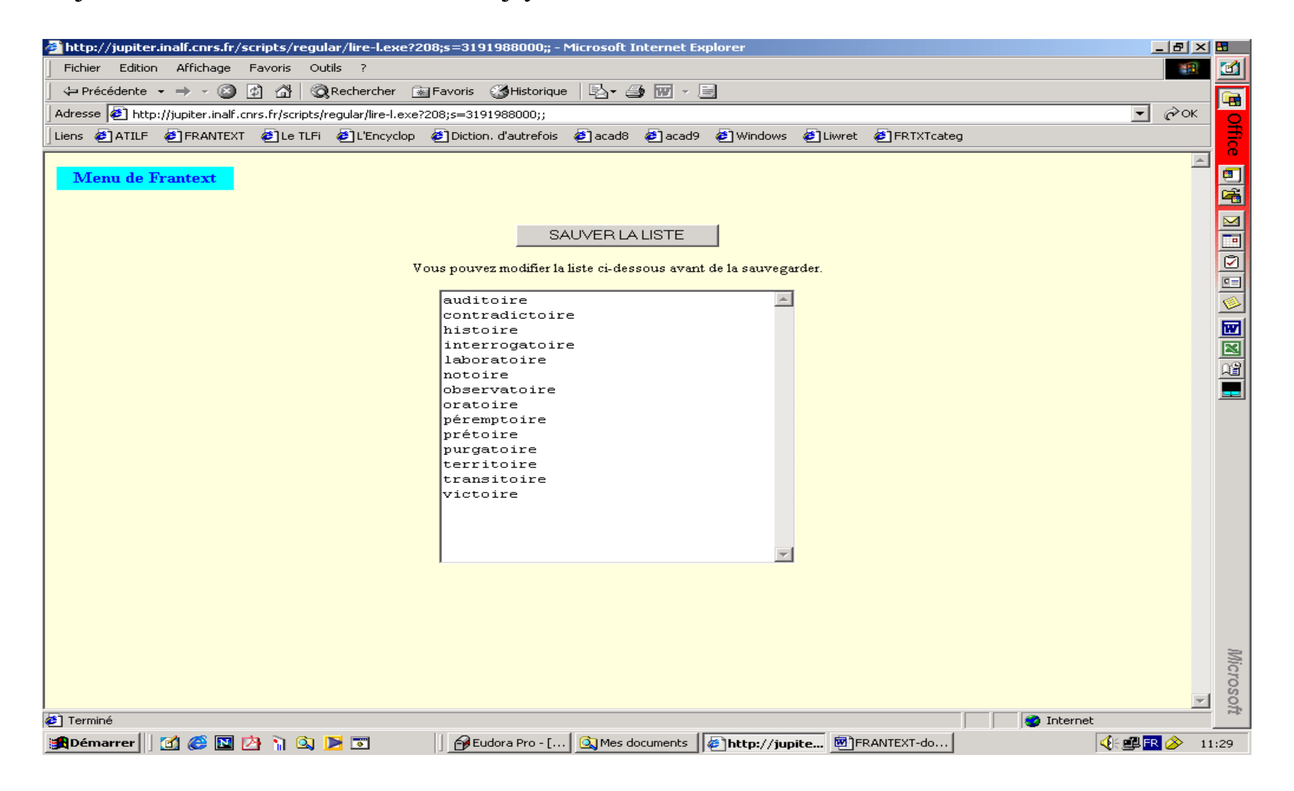
La syntaxe générale d'un critère est la suivante :

- Le symbole **.** (point) désigne n'importe quel caractère.
  Ex. : **.oule** sélectionnera les mots *boule, coule, foule,* etc.
- Une expression telle que [abcd] désigne un caractère qui est soit a, soit b, soit c, soit d.
  Ex. : **coule[sr]** sélectionnera les mots *coules* ou *couler*.
- Une expression telle que [^abcd] désigne un caractère quelconque, à condition qu'il soit différent de a ,b,c ou d.
  Ex. : **[^cr]oule** sélectionnera les mots *foule, houle, poule,* etc. mais pas les mots *coule* et *roule*.

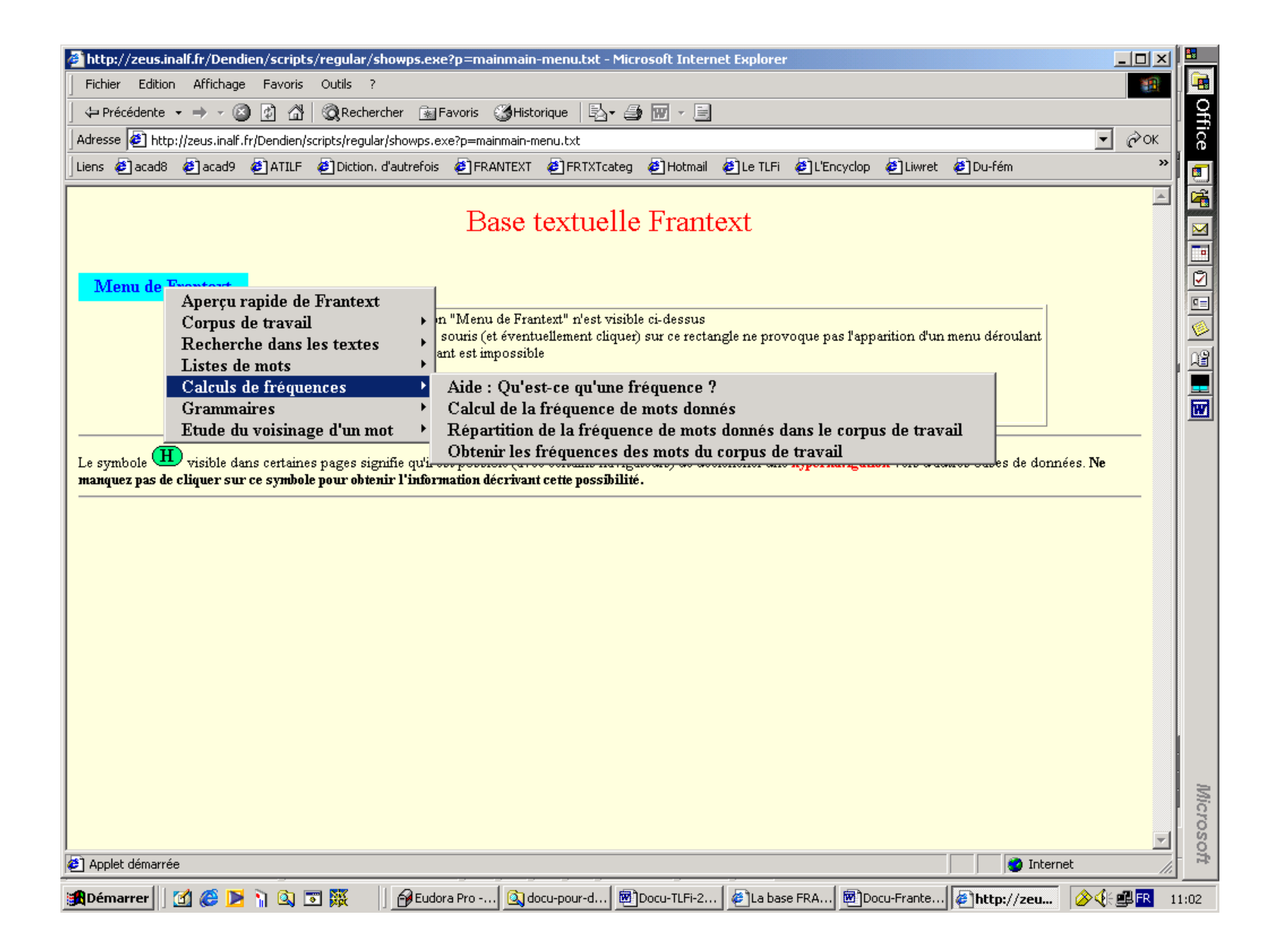
- Les parenthèses servent à délimiter un **groupe** de caractères. Un groupe peut être subdivisé en plusieurs sous-groupes par le symbole |. Ce symbole signifie qu'il y a alternative.
  Ex. : **cheva(l|ux)** sélectionnera *cheval* ou *chevaux*.
  Les parenthèses peuvent être imbriquées à plusieurs niveaux :
  Ex. **(crainti(f|ve)|peureu(x|se))** sélectionnera *craintif, craintive, peureux, peureuse.*
- Symboles de modification.
  - Le symbole * placé derrière un caractère ou un groupe signifie que ce caractère ou groupe peut être absent ou se répéter un nombre quelconque de fois. Ce symbole peut se placer derrière n'importe quel caractère ou groupe, mais il est en fait très pratique lorsqu'il est associé au symbole **.** (point) désignant n'importe quel caractère.
    Ex. : **.*form.*** sélectionnera les mots contenant la chaîne "form" tels que *formée, réforme, informel* etc.
    Ex. : **bla(bla)*** sélectionnera *bla, blabla, blablabla, ...*
  - Le symbole + placé derrière un caractère ou un groupe signifie que ce caractère ou groupe peut se répéter un nombre quelconque de fois (mais au moins une fois).
    Ex. : **.+form.*** sélectionnera les mots contenant la chaîne "form" précédée d'au moins un caractère.
    Ex. : **bla(bla)+** sélectionnera *blabla, blablabla, ..*
  - Le symbole ? placé derrière un caractère ou un groupe signifie que ce caractère ou groupe est optionnel.
    Ex. : **a?politiques?** sélectionnera les mots *politique, politiques, apolitique et apolitiques.*
    Ex. **cinéma(tograph(e|ique))?s?** sélectionnera : *cinéma, cinémas, cinématographe, cinématographes, cinématographique, cinématographiques*.

5. Relecture/modification des listes existantes

Si je choisis de relire la liste Truc3, j'y vois :



Calculs de fréquences



1. Aide : Qu’est ce qu’une fréquence ?

La **fréquence absolue** d’une forme graphique (nous dirons plus simplement « mot ») dans un corpus est le nombre d’occurrences de cette forme dans le corpus.
La **fréquence relative** d’une forme graphique dans un corpus est égale à al fréquence absolue de cette forme divisée par la somme des fréquences absolues de toutes les formes graphiques du corpus.
Ainsi, si le mot *maison* a deux occurrences dans un corpus contenant un million d’occurrences, sa fréquence relative est de deux millionièmes.
Par défaut de qualification, nous utiliserons le terme **fréquence** pour désigner la **fréquence absolue.**

Frantext propose un certain nombre de services capables de calculer des fréquences :

- Calcul des fréquences d’un mot donné, ou des mots d’une liste de mots donnée (dans ce cas la fréquence est donnée pour chaque mot de la liste, et la fréquence cumulée de tous les mots de la liste est calculée). Si vous choisissez cette dernière possibilité, il

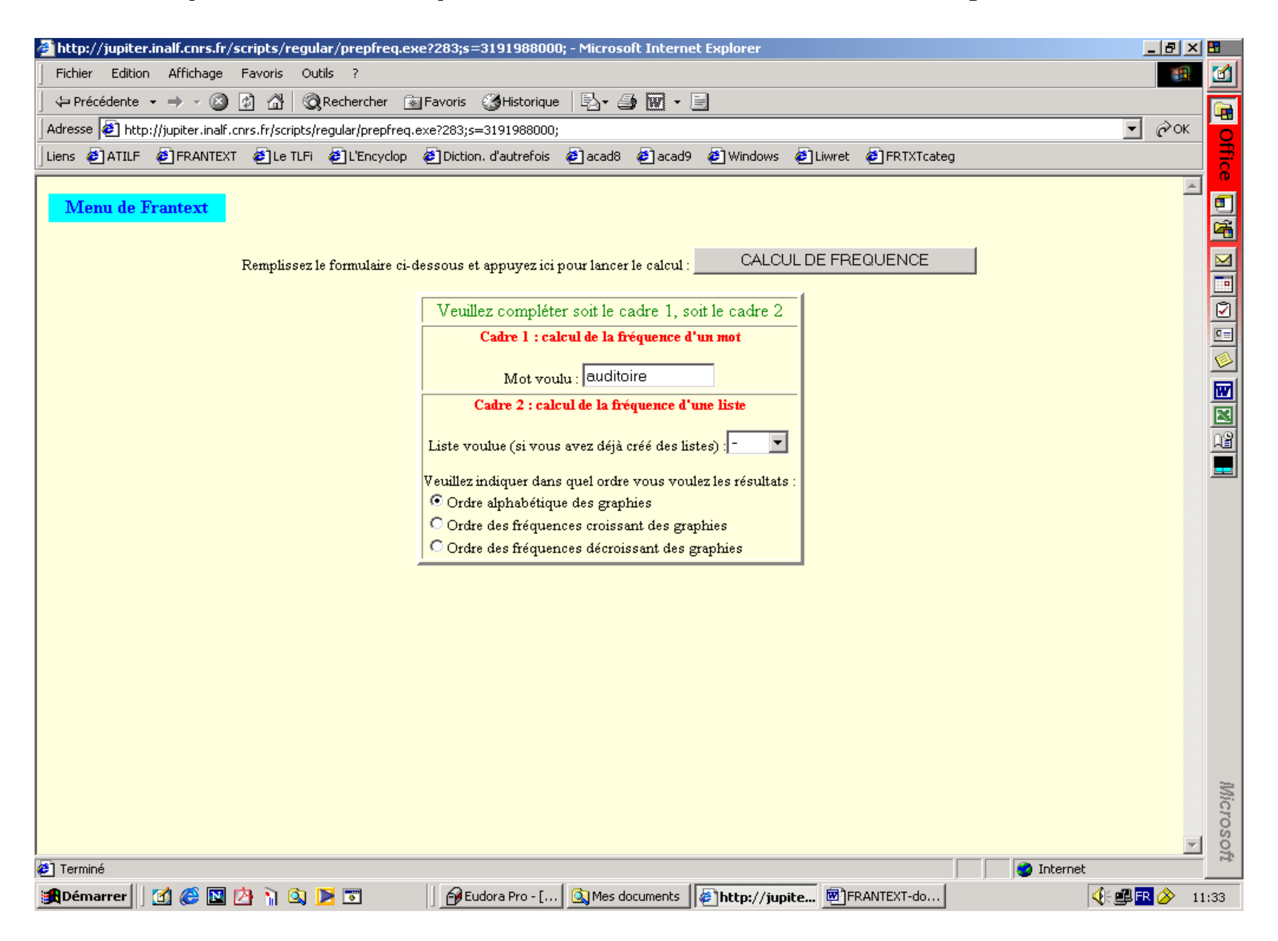
vous faudra fabriquer la liste de mots **avant** de lancer le service de calcul de fréquences.

- Calcul des fréquences des mots du corpus ayant un *profil* donné (par exemple mots commençant par *volo*, ou se terminant par *isme* ou *ismes*).
- Répartition de la fréquence (absolue ou relative) de mots donnés dans les différentes tranches du corpus de travail (le corpus de travail peut être, au choix, découpé en tranches chronologiques, ou en tranches correspondant à l'œuvre de chaque auteur, ou découpé œuvre par œuvre).
- Calcul de la fréquence des mots apparaissant « au voisinage » des occurrences d'un mot donné (par exemple calcul des fréquences des mots apparaissant dans les phrases contenant le mot *abeille*).

Ces services sont facilement accessibles depuis le menu principal de Frantext.

2. Calcul de la fréquence de mots donnés

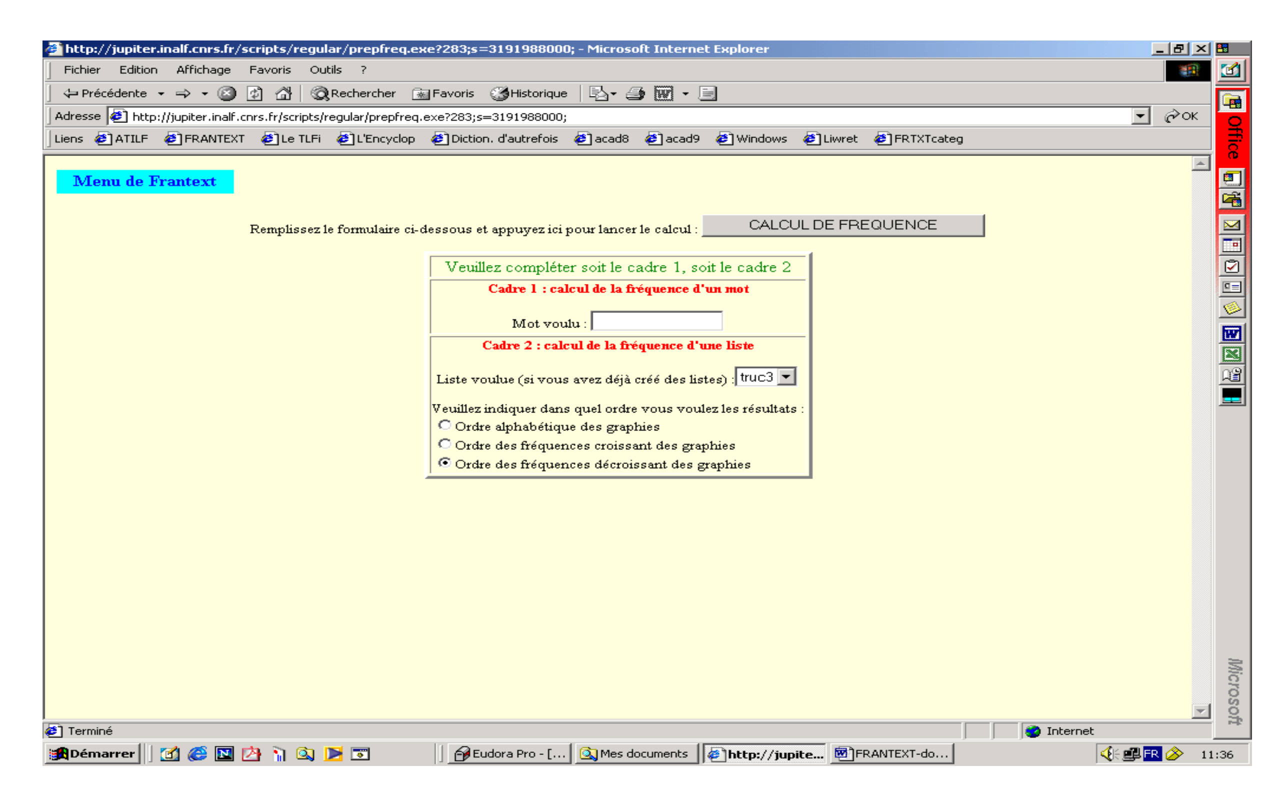
- **Requête sur la fréquence d'un mot donné**, dans le corpus « Leroux » :



Ce qui amène la réponse :

Résultat du calcul de fréquences : auditoire 4

- **Requête, sur la fréquence d'une liste** (le liste truc3), toujours dans le corpus « Leroux » :

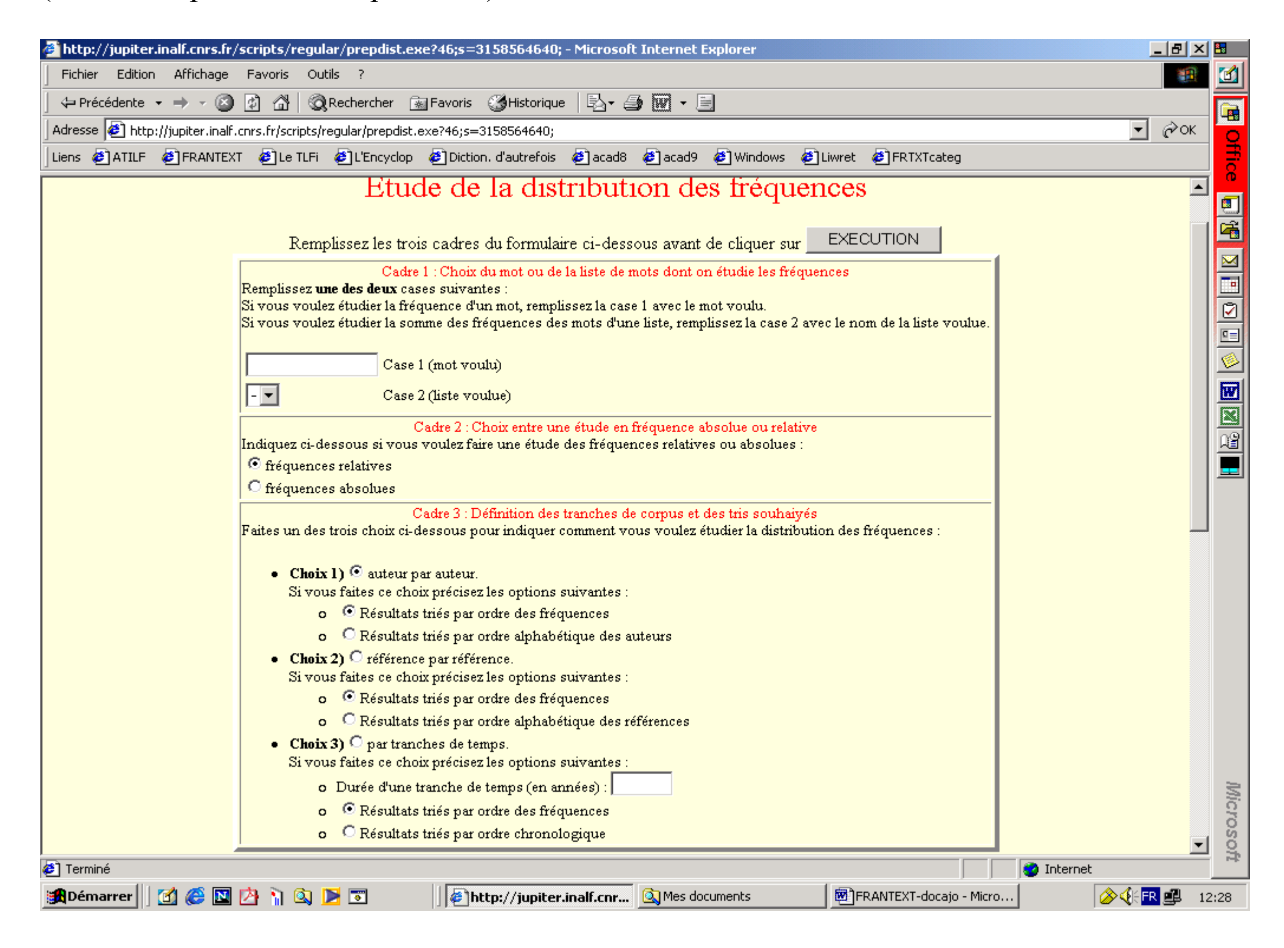


et son résultat :



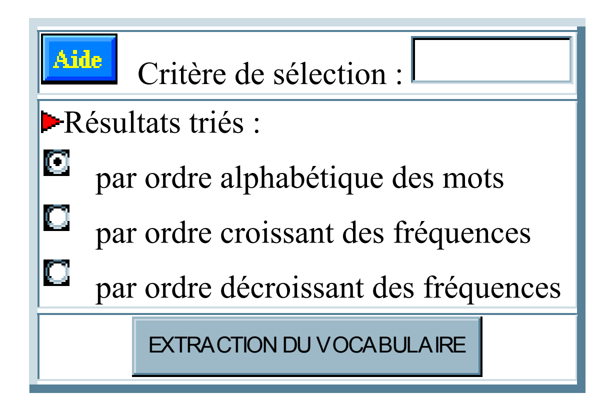
3. Répartition de la fréquence de mots donnés

(dans le corpus de travail prédéfini)

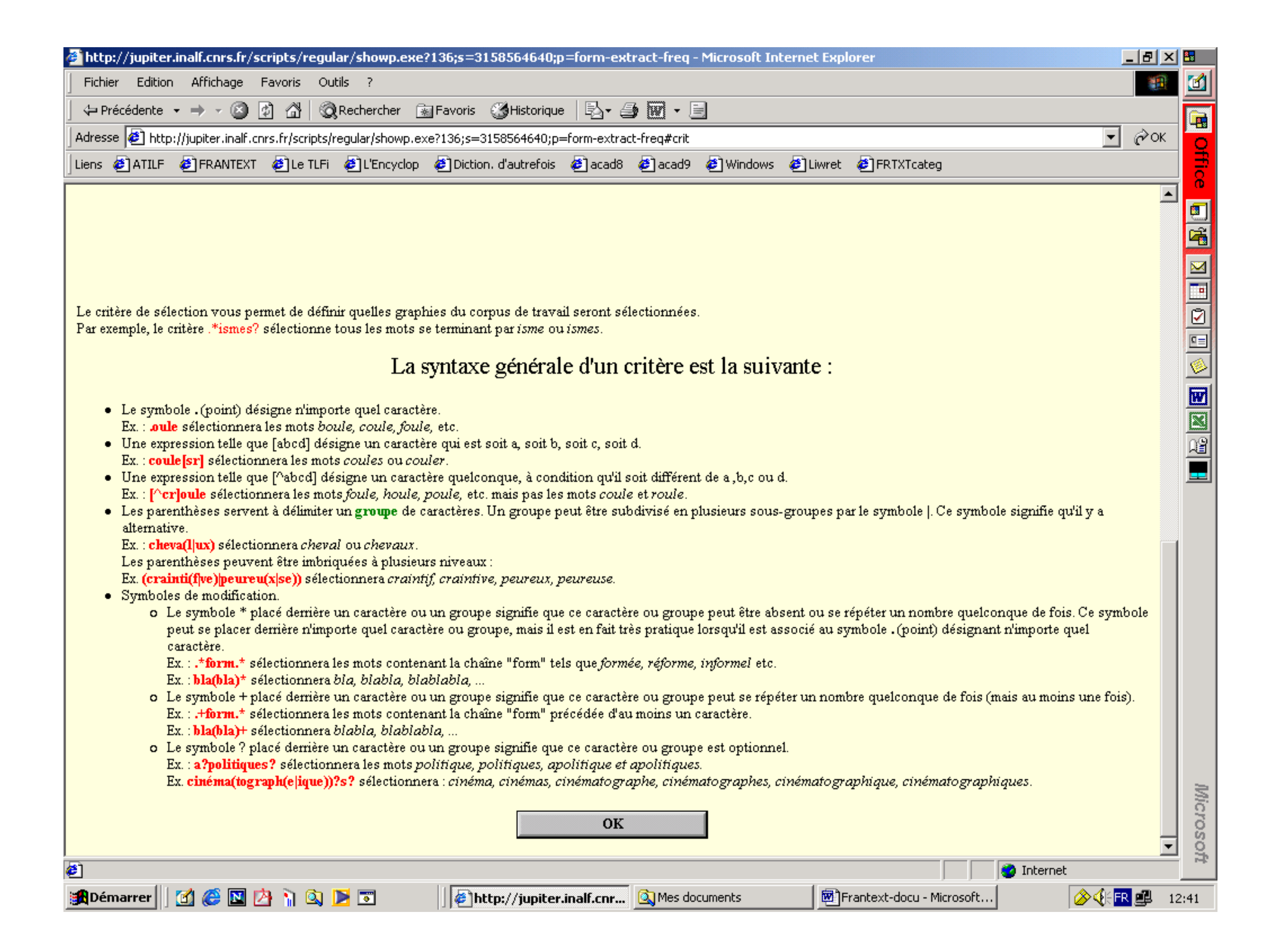


4. Obtenir les fréquences des mots du corpus de travail

Cette requête ouvre sur une fenêtre :



Cette requête renvoie à une aide concernant les critères de sélection (déjà mentionnés précédemment pour les « Listes de mots »):


Le critère de sélection vous permet de définir quelles graphies du corpus de travail seront sélectionnées.
Par exemple, le critère .*ismes? sélectionne tous les mots se terminant par *isme* ou *ismes*.

La syntaxe générale d'un critère est la suivante :

- Le symbole **.** (point) désigne n'importe quel caractère.
  Ex. : **.oule** sélectionnera les mots *boule, coule, foule,* etc.
- Une expression telle que [abcd] désigne un caractère qui est soit a, soit b, soit c, soit d.
  Ex. : **coule[sr]** sélectionnera les mots *coules* ou *couler*.
- Une expression telle que [^abcd] désigne un caractère quelconque, à condition qu'il soit différent de a ,b,c ou d.
  Ex. : **[^cr]oule** sélectionnera les mots *foule, houle, poule,* etc. mais pas les mots *coule* et *roule*.
- Les parenthèses servent à délimiter un **groupe** de caractères. Un groupe peut être subdivisé en plusieurs sous-groupes par le symbole |. Ce symbole signifie qu'il y a alternative.
  Ex. : **cheva(l|ux)** sélectionnera *cheval* ou *chevaux*.
  Les parenthèses peuvent être imbriquées à plusieurs niveaux :
  Ex. **(crainti(f|ve)|peureu(x|se))** sélectionnera *craintif, craintive, peureux, peureuse.*
- Symboles de modification.
  - Le symbole * placé derrière un caractère ou un groupe signifie que ce caractère ou groupe peut être absent ou se répéter un nombre quelconque de fois. Ce symbole peut se placer derrière n'importe quel caractère ou groupe, mais il est en fait très pratique lorsqu'il est associé au symbole **.** (point) désignant n'importe quel caractère.
    Ex. : **.*form.*** sélectionnera les mots contenant la chaîne "form" tels que *formée, réforme, informel* etc.
    Ex. : **bla(bla)*** sélectionnera *bla, blabla, blablabla, ...*
  - Le symbole + placé derrière un caractère ou un groupe signifie que ce caractère ou groupe peut se répéter un nombre quelconque de fois (mais au moins une fois).
    Ex. : **.+form.*** sélectionnera les mots contenant la chaîne "form" précédée d'au moins un caractère.
    Ex. : **bla(bla)+** sélectionnera *blabla, blablabla, ...*
  - Le symbole ? placé derrière un caractère ou un groupe signifie que ce caractère ou groupe est optionnel.
    Ex. : **a?politiques?** sélectionnera les mots *politique, politiques, apolitique et apolitiques.*
    Ex. **cinéma(tograph(e|ique))?s?** sélectionnera : *cinéma, cinémas, cinématographe, cinématographes, cinématographique, cinématographiques.*

OK

<u>En plus lisible cela donne</u> :

Le critère de sélection vous permet de définir quelles graphies du corpus de travail seront sélectionnées.
Par exemple, le critère .*ismes? sélectionne tous les mots se terminant par *isme* ou *ismes*.

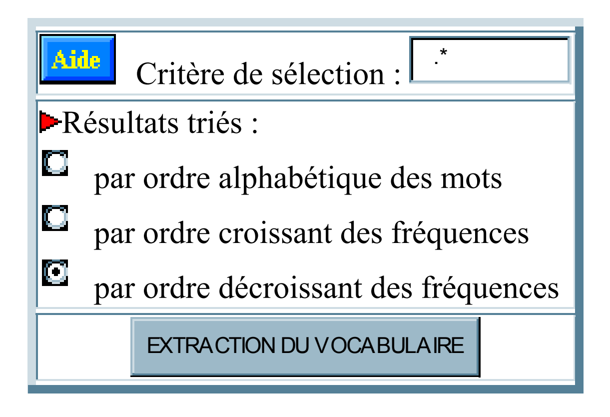
La syntaxe générale d'un critère est la suivante :

- Le symbole **.** (point) désigne n'importe quel caractère.
  Ex. : **.oule** sélectionnera les mots *boule, coule, foule,* etc.
- Une expression telle que [abcd] désigne un caractère qui est soit a, soit b, soit c, soit d.
  Ex. : **coule[sr]** sélectionnera les mots *coules* ou *couler*.
- Une expression telle que [^abcd] désigne un caractère quelconque, à condition qu'il soit différent de a ,b,c ou d.
  Ex. : **[^cr]oule** sélectionnera les mots *foule, houle, poule,* etc. mais pas les mots *coule* et *roule*.
- Les parenthèses servent à délimiter un **groupe** de caractères. Un groupe peut être subdivisé en plusieurs sous-groupes par le symbole |. Ce symbole signifie qu'il y a alternative.
  Ex. : **cheva(l|ux)** sélectionnera *cheval* ou *chevaux*.
  Les parenthèses peuvent être imbriquées à plusieurs niveaux :
  Ex. **(crainti(f|ve)|peureu(x|se))** sélectionnera *craintif, craintive, peureux, peureuse.*
- Symboles de modification.
  - Le symbole * placé derrière un caractère ou un groupe signifie que ce caractère ou groupe peut être absent ou se répéter un nombre quelconque de fois. Ce symbole peut

se placer derrière n'importe quel caractère ou groupe, mais il est en fait très pratique lorsqu'il est associé au symbole **.** (point) désignant n'importe quel caractère.
Ex. : **.*form.*** sélectionnera les mots contenant la chaîne "form" tels que *formée, réforme, informel* etc.
Ex. : **bla(bla)*** sélectionnera *bla, blabla, blablabla, ...*

- Le symbole + placé derrière un caractère ou un groupe signifie que ce caractère ou groupe peut se répéter un nombre quelconque de fois (mais au moins une fois).
  Ex. : **.+form.*** sélectionnera les mots contenant la chaîne "form" précédée d'au moins un caractère.
  Ex. : **bla(bla)+** sélectionnera *blabla, blablabla, ...*
- Le symbole ? placé derrière un caractère ou un groupe signifie que ce caractère ou groupe est **ma(tograph(e|ique))?s?** sélectionnera : *cinéma, cinémas, cinématographe, cinématographes, cinématographique, cinématographiques*optionnel.
  Ex. : **a?politiques?** sélectionnera les mots *politique, politiques, apolitique et apolitiques.*
  Ex. **ciné**

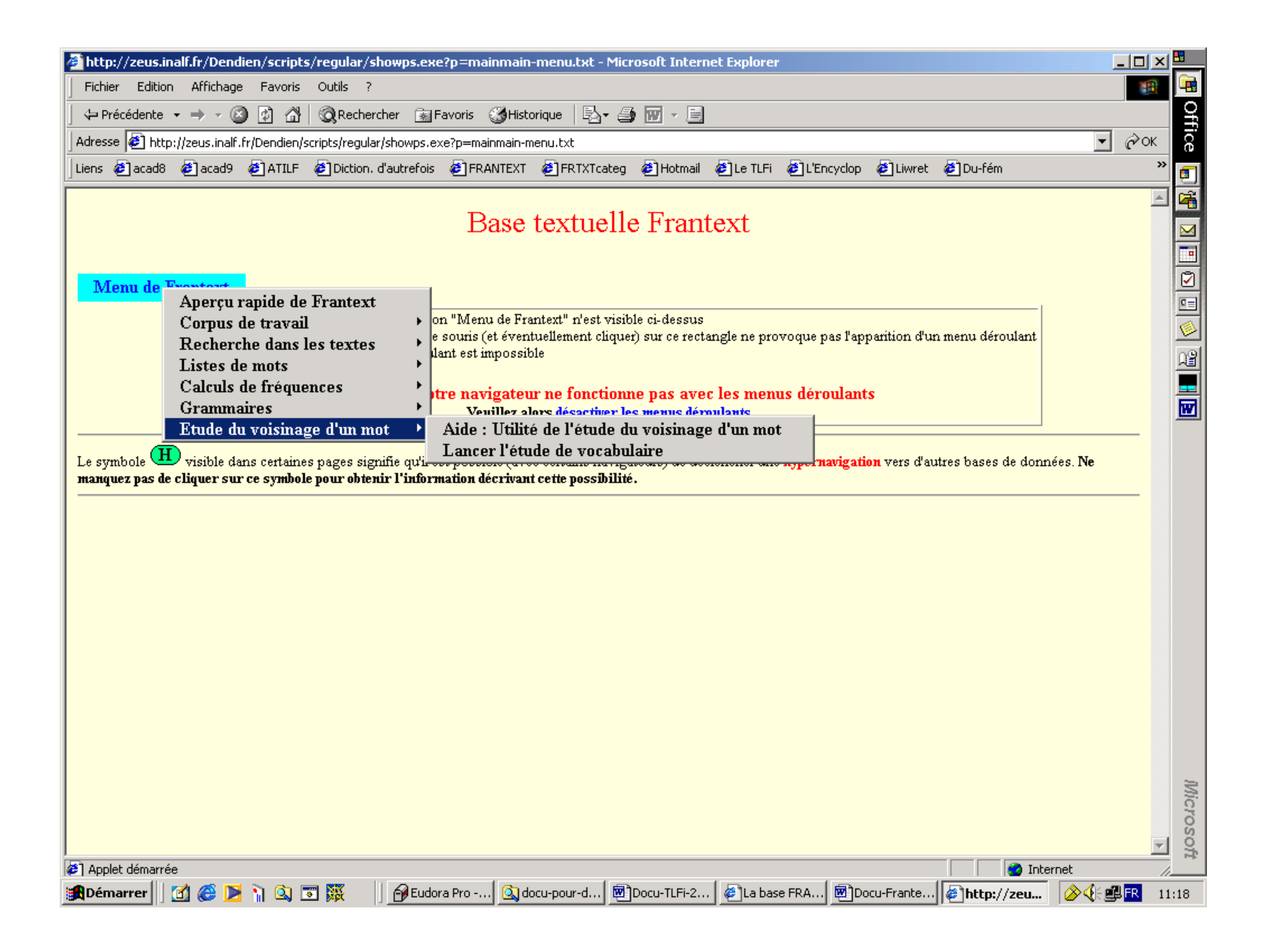
Si par exemple, je souhaite une liste de tous les mots de mon corpus, par ordre de fréquence décroissante, je demanderai :



Je saurai ainsi que, dans le corpus Leroux, il y a :

41487 ,
20587 de
14901 la
12691 et
11569 le
10932 -
10621 l'
9444 à
9293 il
8427 que
6509 est
6036 les
etc.

Grammaires



1. Aide : A quoi servent les grammaires ?

1.a) Des recherches simples pas si simples que ça

Supposons que l'on se propose de rechercher dans les textes des emplois de l'expression **prendre la fuite**.
Une recherche, apparemment aussi simple que celle-ci, pose en fait un problème complexe. En effet, une occurrence de cette expression peut revêtir des formes très différentes : ***prenait la fuite, prenait toujours la fuite, ne prendront pas la fuite, avions pris la fuite, n'avions pas pris la fuite, prenez-vous la fuite, ne prendrions-nous pas volontiers la fuite, a-t-il pris la fuite, n'a-t-il jamais pris la fuite,*** etc.
Le plus souvent, c'est la syntaxe du français qui est à l'origine de la multiplicité des formes de l'expression recherchée. Dans d'autres cas, la multiplicité est due au fait que l'utilisateur recherche un objet textuel pouvant par essence revêtir des formes très variables : par exemple une indication de période peut se manifester sous une des formes suivantes :

- 1939-1945
- 1939-45
- janvier-mars 1856
- janvier 1855 - mars 1856
- 15 janvier - 11 mars 1856
- 5-10 juillet 1711
- 5 janvier 1903 - 6 juin 1904

Quelle que soit la raison conduisant à la multiplicité des formes de l'objet recherché, Frantext vous propose un moyen d'exprimer de manière formelle, simple et structurée l'ensemble de ces formes : **les grammaires**.
Pourquoi le terme de "grammaire" ? Le mot **grammaire** est ici employé dans le sens qu'il revêt en Théorie de Langages, et non pas dans le sens "grammaire d'une langue naturelle". C'est à dire que les grammaires que vous allez écrire sont capables d'engendrer, par la combinaison de leurs règles, un certain "langage" dont les occurrences seront recherchées par Frantext dans les textes que vous étudiez.

1.b) Un exemple de grammaire

L'exemple ci-dessous donne une idée de ce à quoi ressemble une grammaire (nous y reprenons l'exemple des périodes).

mois :
 janvier|février|mars|avril|mai|juin|juillet|août|septembre|octobre|novembre|décembre
chiffre :
 0|1|2|3|4|5|6|7|8|9
annee :
 1 **&r**chiffre **&r**chiffre **&r**chiffre
annee_a_annee :
 &rannee - (**&r**annee | **&r**chiffre **&r**chiffre)
mois_a_mois :
 &rmois - **&r**mois **&r**annee |
 &rmois **&r**annee - **&r**mois **&r**annee
jour :
 1 er | **&r**chiffre &? **&r**chiffre
jour_a_jour :
 &rjour - **&r**jour **&r**mois **&r**annee |
 &rjour **&r**mois - **&r**jour **&r**mois **&r**annee|
 &rjour **&r**mois **&r**annee - **&r**jour **&r**mois **&r**annee
periode :
 &ranne_a_annee|**&r**mois_a_mois|**&r**jour_a_jour

Cette grammaire définit un certain nombre de règles (dont les noms sont indiqués en bleu):

- **mois** : liste des différents mois.
- **chiffre** : liste des différents chiffres.
- **annee** : définit l'"année" comme étant le chiffre 1 suivi de trois fois l'application de le règle "chiffre", c'est à dire comme un 1 suivi de trois chiffres. On remarquera la mention "&r" (indiquée en rouge dans la grammaire pour la commodité visuelle) qui signifie que l'on **applique** la règle chiffre.

- **annee_a_annee** : définit "annee_a_annee" comme étant une année suivie d'un tiret suivi soit d'une année, soit de deux chiffres. La règle "annee_a_annee" est donc capable d'engendrer des contextes tels que ***1939-1945*** ou ***1939-45***.
- **mois_a_mois** : définit "mois_a_mois" soit comme deux noms de mois séparés par un tiret et suivis d'une indication d'annéee (la règle "mois_a_mois" est alors capable d'engendrer des contextes tels que ***janvier-mars 1856***), soit comme un nom de mois suivi d'une année, suivie d'un tiret, suivi d'un nom de mois suivi d'une année (la règle "mois_a_mois" est alors capable d'engendrer des contextes tels que ***janvier 1855 - mars 1856***).
- **jour** : définit "jour" comme étant l'indication **1er** (pour premier) ou un chiffre éventuellement (l'éventualité est notée &? dans la grammaire) suivi d'un autre chiffre.
- **jour_a_jour** : définit "jour_jour" sous trois formes possibles, capables d"engendrer respectivement des contextes tels que ***5-10 juillet 1711***, ***15 janvier - 11 mars 1856*** et ***5 janvier 1903 - 6 juin 1904***.
  NOTA : en faisant jouer la factorisation, on peut ramener la regle "jour_a_jour" a la forme suivante, beaucoup plus simple :
  &rjour &?(**&r**mois &?**&r**annee) -**&r**jour **&r**mois **&r**annee
- **periode** : définit "période" comme étant au choix une des règles une "annee_a_annee", soit une "mois_a_mois", soit une "jour_a_jour". Autrement dit, la règle "periode" peut engendrer toutes les possibilités d'expressions de fourchette de dates que nous avions envisagées au départ.

1.c) Comment créer et utiliser une grammaire.

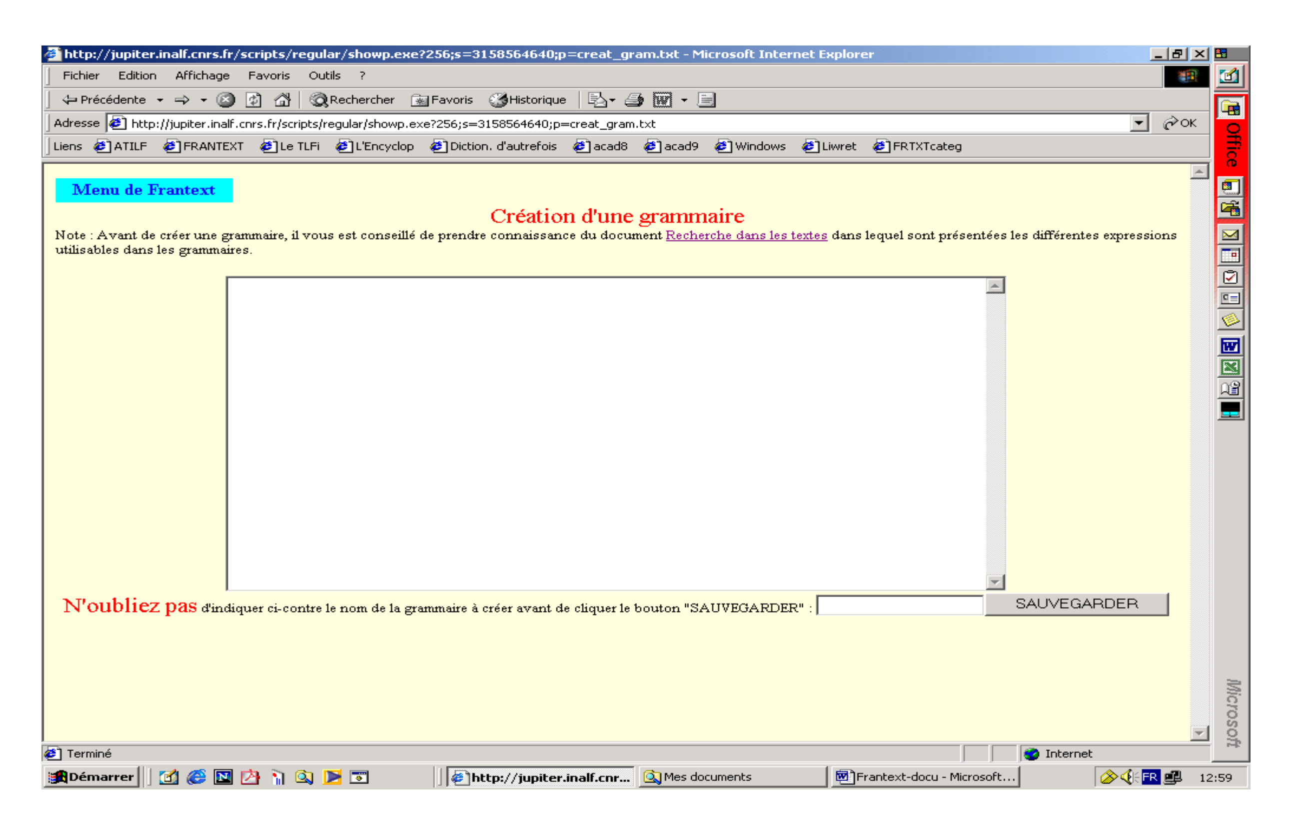
- Création, modification, téléchargement.
  - Depuis le "menu principal" de Frantext, vous avez accès au service "création de grammaire" qui vous ouvre une espace pour la saisie de la grammaire. A l'issue de cette saisie, vous pourrez sauvegarder la grammaire dans un fichier. Cette grammaire existe alors **sur notre serveur Internet**.
  - Si votre grammaire comporte des erreurs, vous pourrez la relire (service "relecture/modification/téléchargement") pour la corriger ou la compléter.
  - Si votre grammaire est "au point", vous pouvez la télécharger depuis notre serveur pour la sauvegarder sur votre ordinateur. Pour pourrez, lors de sessions de travail futures, télécharger votre grammaire de votre ordinateur vers notre serveur et la réutiliser pour vos recherches.
- Utilisation de la grammaire dans des recherches textuelles.
  Supposons que la grammaire définissant les différentes périodes de temps ait été ait été écrite dans un fichier de nom "**epoque**".
  Supposons que l'on veuille rechercher des attestations de périodes de type 1939-45 (année à année). L'expression de recherche (voir le service "recherche dans les textes") que vous devrez taper sera **&rannee_a_annee,epoque**. Cette expression signifie que vous recherchez des contextes engendrés par la règle "annee_a_annee" définie dans la grammaire "epoque".
  Si vous voulez chercher la forme la plus générale des expressions de périodes, et voulez qu'elle soit précédée de "en", vous taperez **en &rperiode,epoque**.

1.d) Conclusion.

Les **grammaires** constituent un moyen extrêmement puissant pour la recherche de contextes complexes qui vous permet de vous attaquer à des classes de problèmes **qu'il serait**

impensable de résoudre avec tout autre système que Frantext.
L'apparente complexité d'une grammaire vue dans sa globalité est compensée par le fait qu'une grammaire s'écrit **règle par règle**, et que chaque règle, si elle est trop complexe, peut s'écrire en écrivant des sous-règles beaucoup plus simples. Une grammaire peut donc être considérée comme un édifice composé d'éléments **extrêmement simples (les règles)** dont l'assemblage produit un résultat pouvant être **extrêmement complexe (la description des contextes recherchés)**. Nous vous conseillons donc de commencer par mettre en oeuvre des "mini-grammaires" très simples (une ou deux règles). Une fois cette étape franchie, vous constaterez que l'écriture de grammaires complexes se fait avec la plus grande facilité.
Vous trouverez dans le document Recherche dans les textes l'arsenal mis à votre disposition pour les recherches textuelles et la réalisation de grammaires. Entre autres choses, vous y découvrirez que les grammaires peuvent être **paramétrées**, ce qui vous donne la possibilité de les utiliser comme macro-instructions de recherche. Vous verrez également que plusieurs grammaires peuvent être utilisées concurremment et se référencer les unes les autres, ce qui vous donne la possibilité de vous doter d'une véritable "bibliothèque de grammaires" adaptée à vos recherches.
Note : Des détails complémentaires sont consultables dans la 2ème section, consacrée à Frantext catégorisé. Un chapitre est consacré à « Entités catégorisées et Grammaires ».

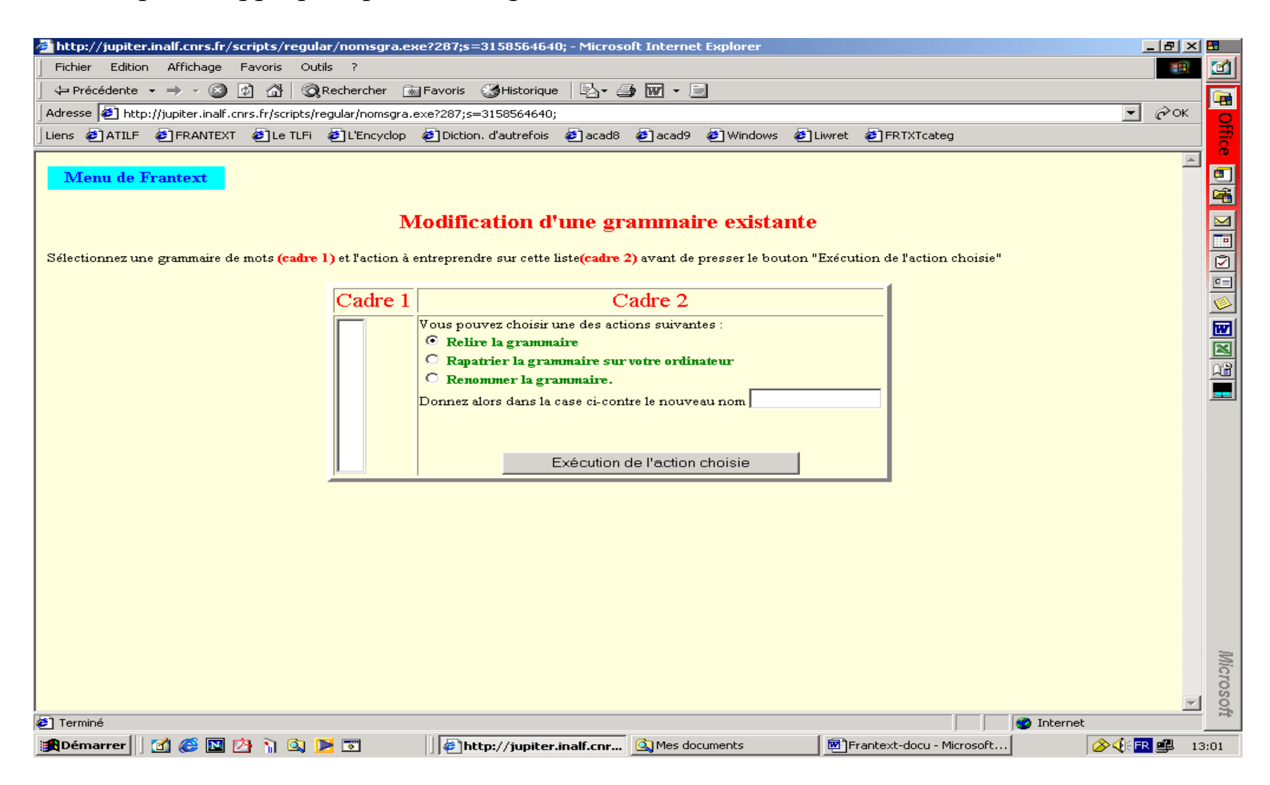
2. Création d'une grammaire



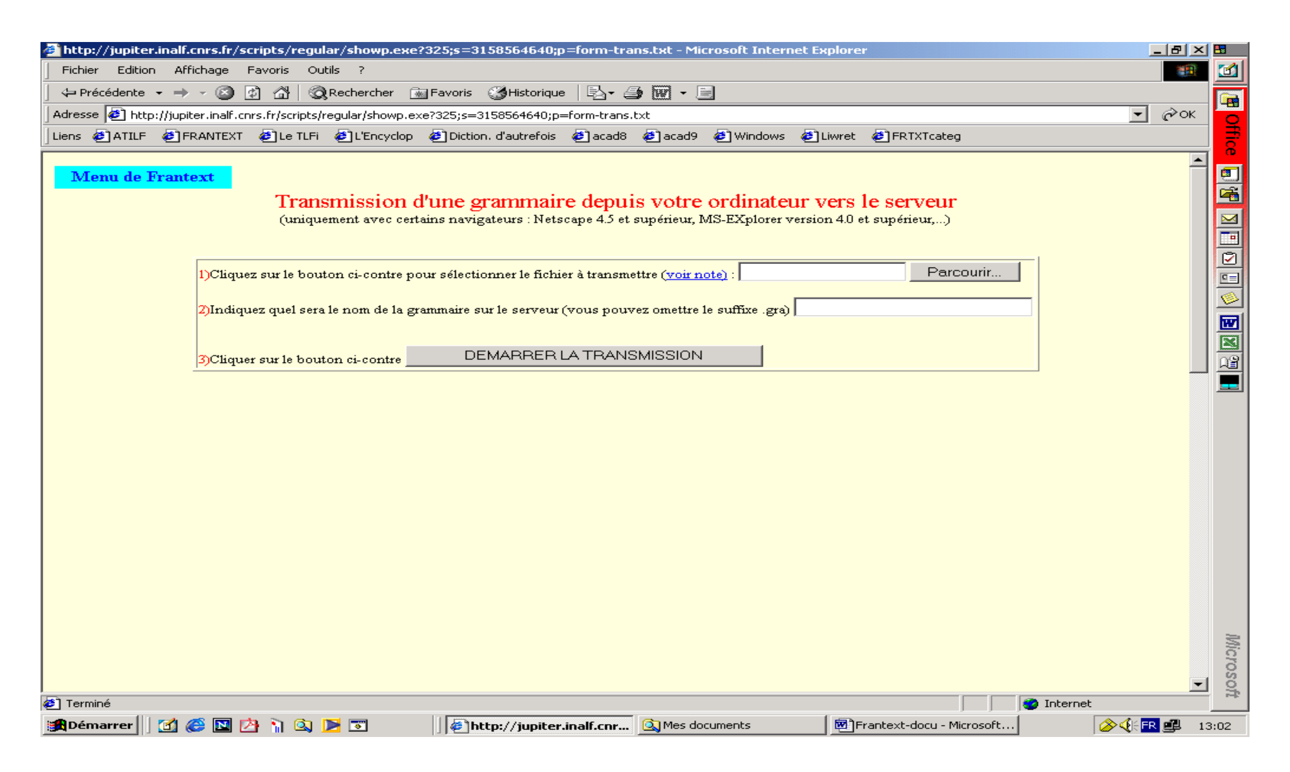
Cette fenêtre renvoie au chapitre Recherche dans les textes
Pour voir un exemple de grammaire et son résultat, aller à « Exemples de requêtes », voir l'exemple 3 (cf. Table des matières).

3. Relecture/modification/téléchargement

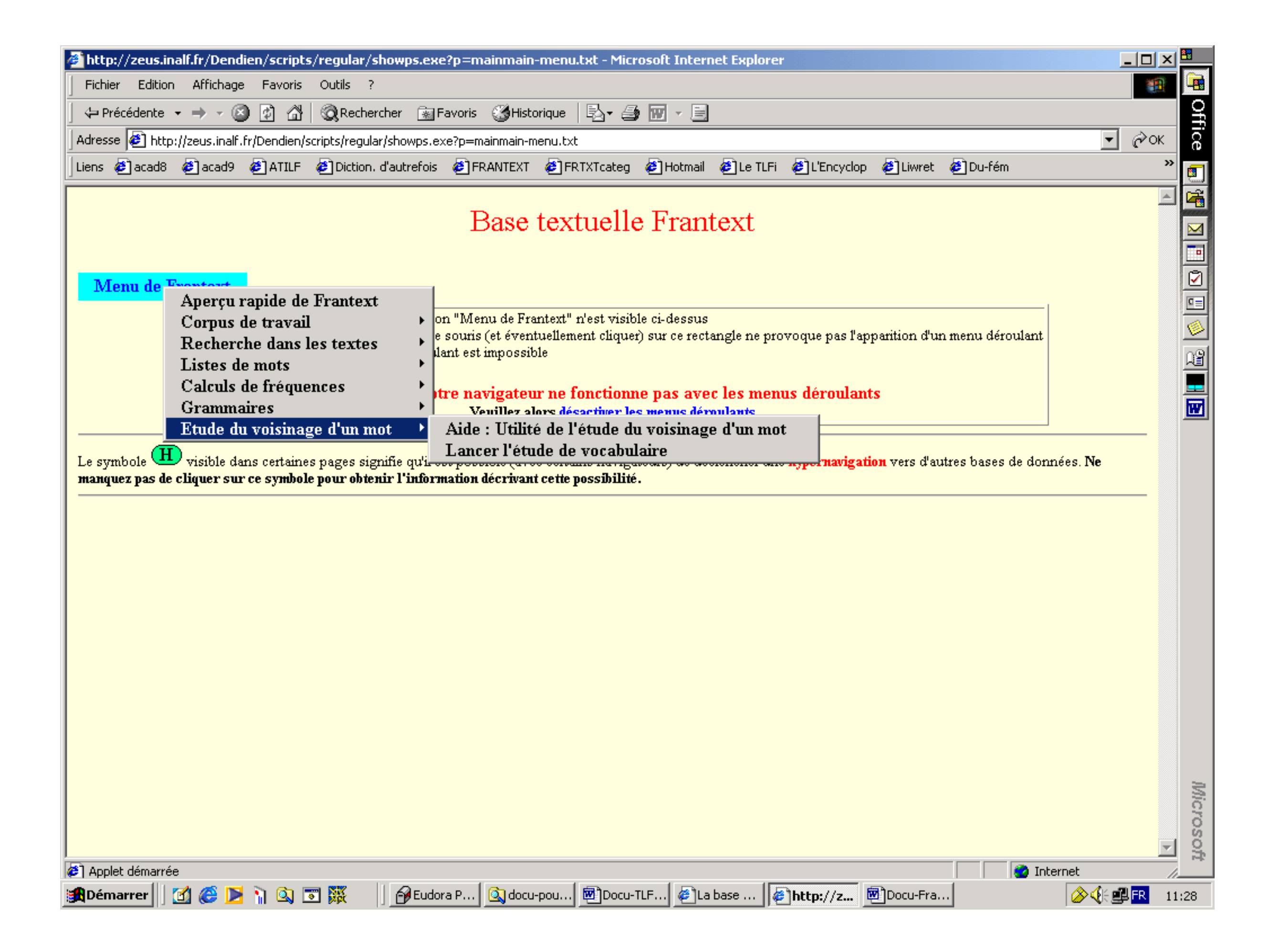
Ceci ne peut s'appliquer que sur des grammaires existantes .



4. Télécharger une grammaire



Étude du voisinage d'un mot



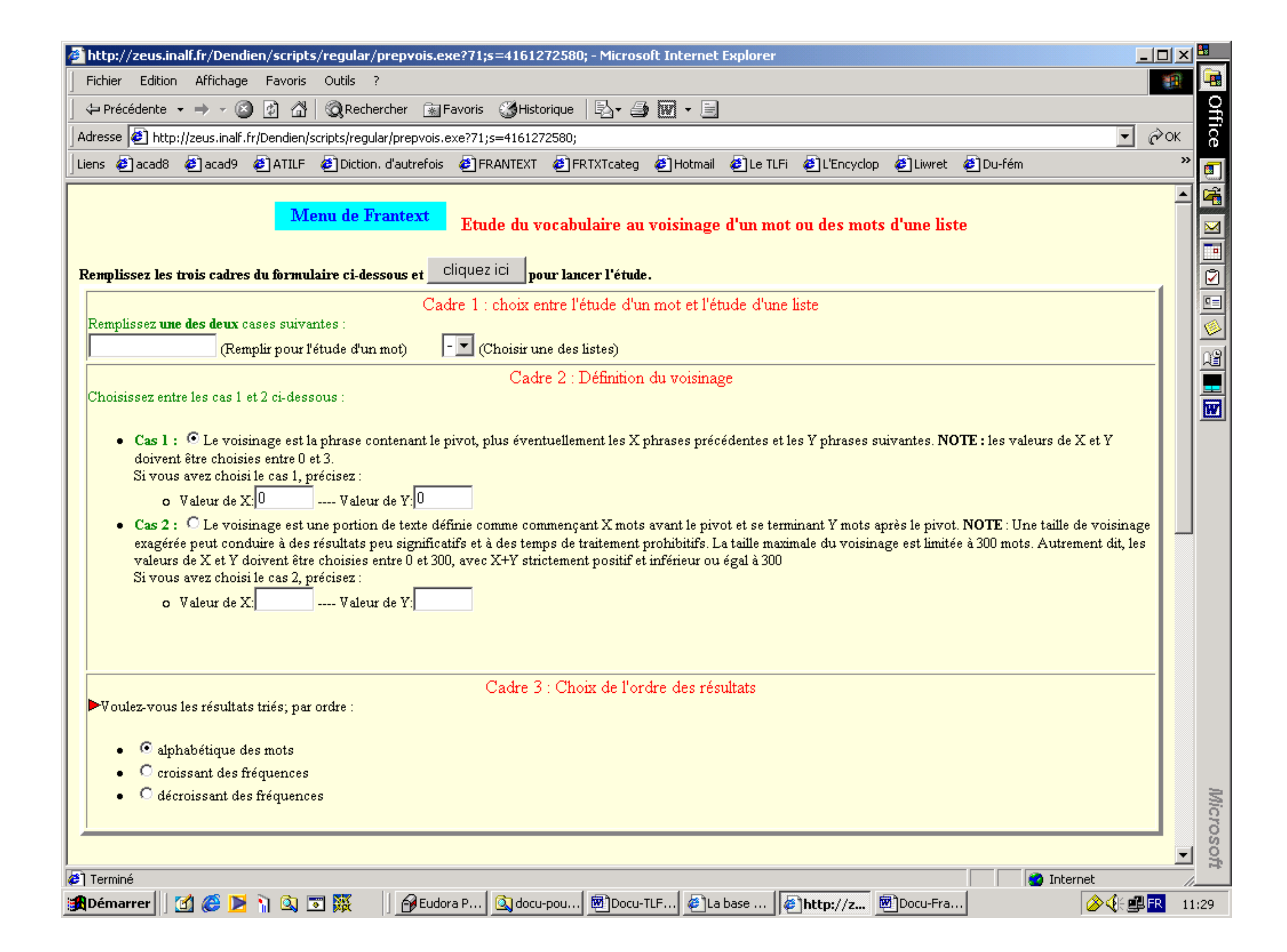
1. Aide : Utilité de l'étude du voisinage d'un mot

Dans certaines études de type thématique, il est intéressant d'étudier les mots qui sont « associés » à un mot donné. Par exemple, partant du mot *abeille* (Appelons **pivot** cette occurrence de mot donné), il peut être intéressant de procéder à une étude systématique des mots utilisés dans les mêmes phrases qu'*abeille* ou utilisés dans un voisinage donné (X mots avant *abeille*, Y mots après *abeilles*).

C'est à une telle étude qu'est consacré le service « Étude du vocabulaire au voisinage des occurrences d'un mot donné ».

Les résultats rendus par ce service sont constitués de la liste des mots trouvés au voisinage du mot donné, triés au choix par ordre alphabétique, ordre croissant ou décroissant des fréquences, et sans précision de position (à droite ou à gauche) par rapport au mot pivot. Il ne s'agit aucunement de concordances, mais de simples indications qui prennent en compte une notion de « proximité » .

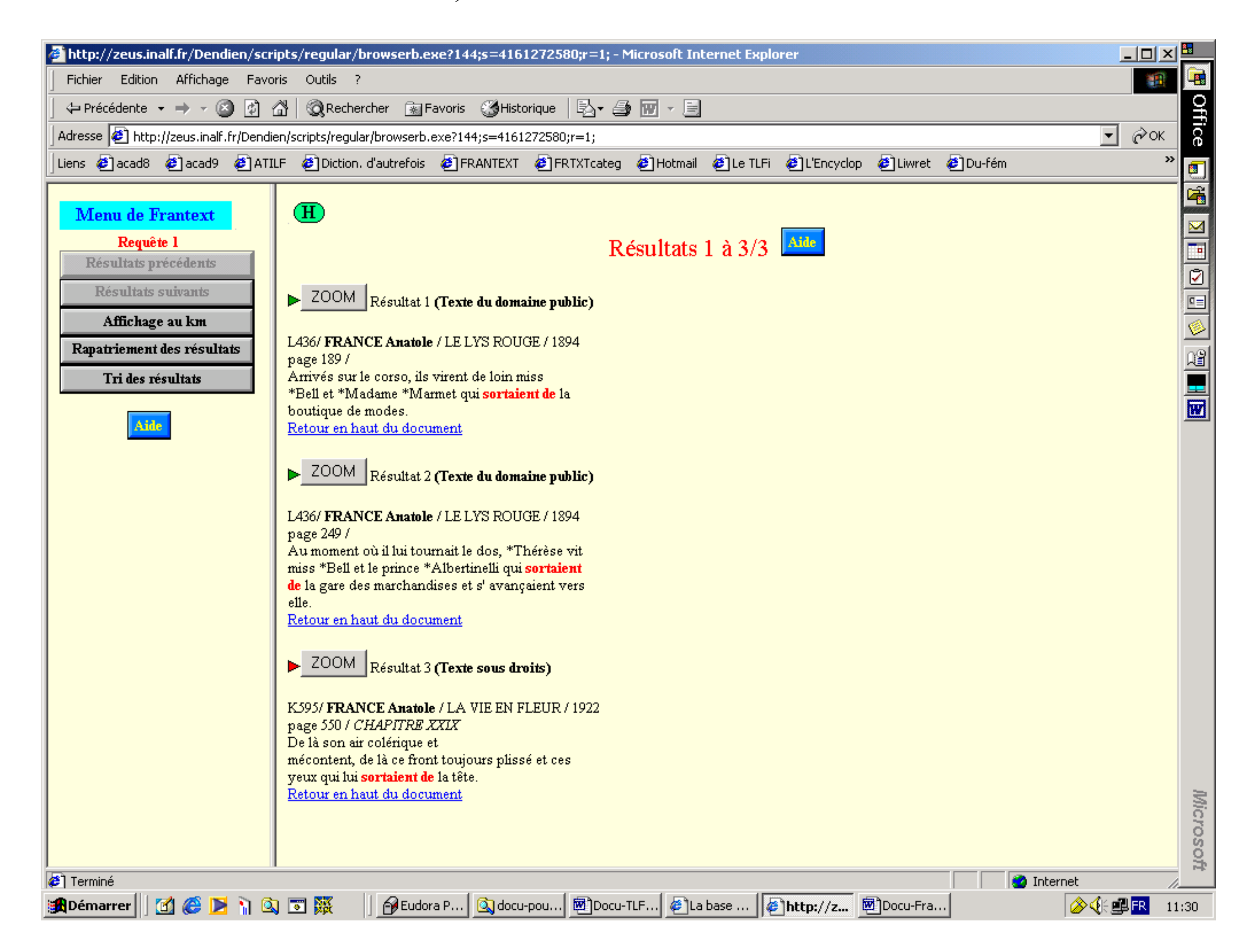
2. Lancer l'étude de vocabulaire



Note, lue quelque part…..
Sur le temps d'exécution :
Ce service met en jeu des moyens de traitement très importants. Il doit donc être utilisé avec mesure.
Pour cette raison, ce programme s'arrête automatiquement lorsque le nombre d'occurrences dépasse **10000**. Dans ce cas, les résultats sont quand même obtenus, mais ne correspondent qu 'au dépouillement des 10000 premiers voisinages.
Il est donc prudent de s'assurer de la fréquence du pivot grâce au service « Calcul de la fréquence d'un mot ou de chaque mot d'une liste ».

Visualisation/Rapatriement des résultats

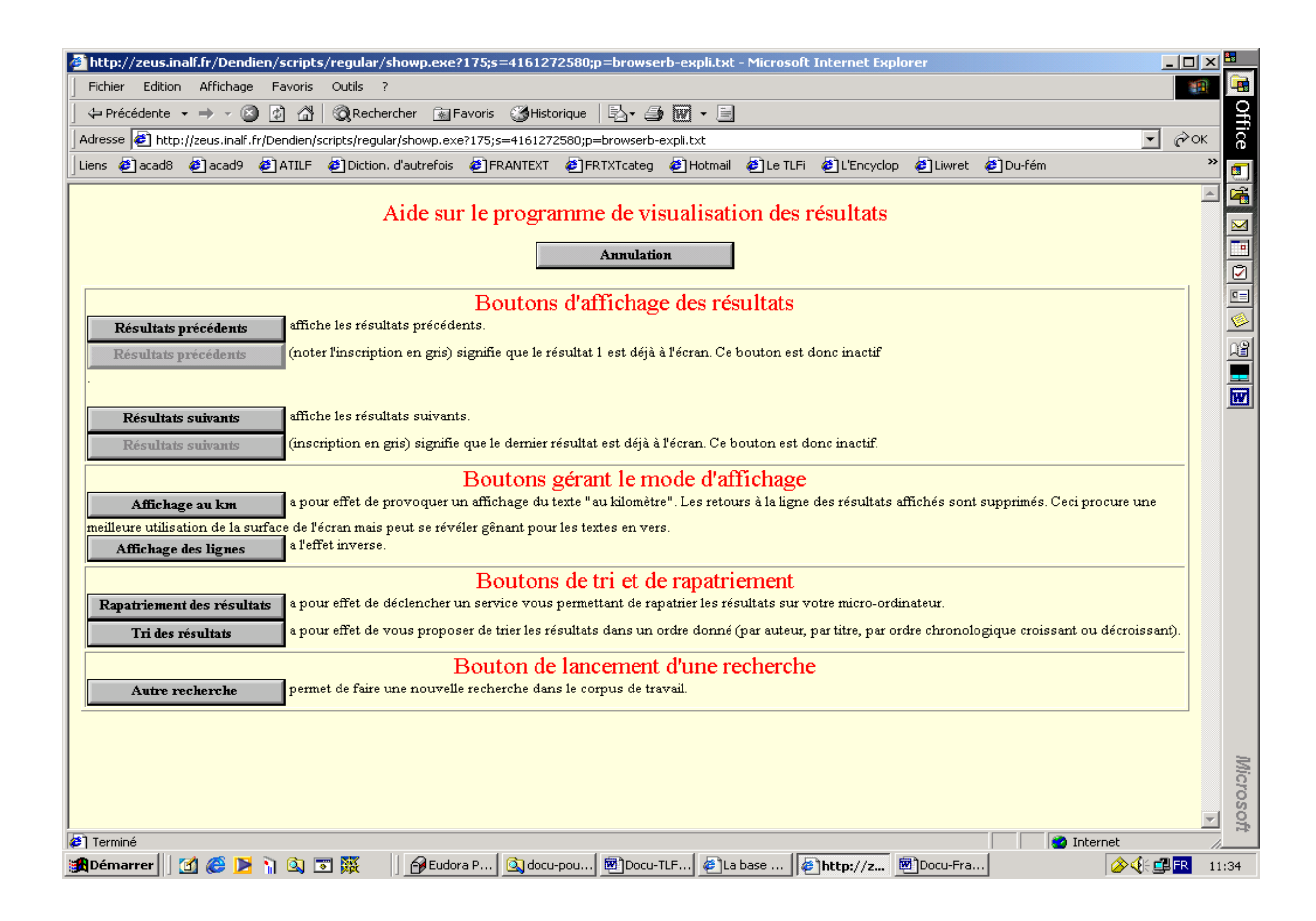
A partir d'une requête portant sur : Auteur = **France**
Séquence recherchée :*sortaient de*
on obtient les 3 résultats suivants, dans la fenêtre ci-dessous :



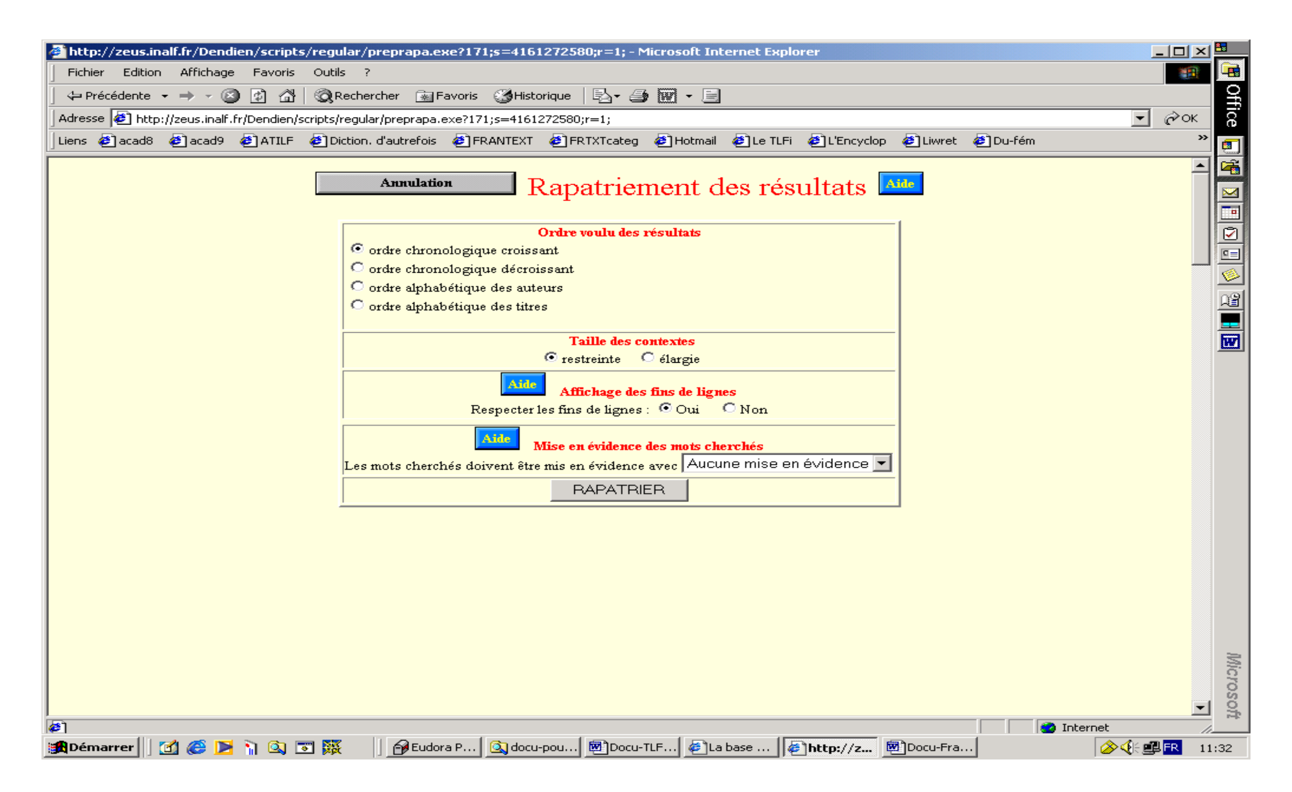
Cette fenêtre propose, dans la zone de gauche, trois choses intéressantes :

- Un bouton d'aide, qui renvoie sur le programme de visualisation des résultats,
- Un bouton proposant le rapatriement des résultats,
- Un bouton proposant le tri des résultats.

1. Le bouton d'aide :



2. Le rapatriement des résultats



Aide : Le rapatriement des résultats :

Le principe de ce service est de vous permettre de récupérer sur votre micro-ordinateur un fichier contenant tous les résultats correspondant à la dernière recherche que vous avez effectuée.
Lorsque vous allez activer ce service, le serveur vous enverra les résultats. Le déroulement exact des opérations va dépendre du navigateur Internet que vous utilisez et **en aucun cas de la manière dont sont réalisés les logiciels de Frantext**. En principe, votre navigateur va vous demander (pas forcément en français si vous n'utilisez pas une version française !) ce que vous voulez faire des informations envoyées par le serveur.

Les choix possibles sont en général :

- Enregistrer dans un fichier disque dont vous pouvez choisir le nom et l'emplacement. Dans ce cas, après réception des résultats, vous pourrez ouvrir ce fichier avec un éditeur de texte standard de votre micro-ordinateur.
- Activer un éditeur de texte qui recevra directement le flux des résultats. Dans ce cas, si votre micro-ordinateur comporte plusieurs éditeurs de texte, le choix de l'éditeur activé dépend de la manière dont votre système et votre navigateur sont configurés **en aucun cas des logiciels de Frantext**.

Note sur l'encodage des résultats

Les résultats sont encodés suivant la norme ISO 8859-1. Cette norme définit la grille de codes des caractères. Elle est compatible avec les usages en vigueur sur les systèmes MS-Windows et UNIX. Elle est compatible avec MacIntosh
Cependant, il se peut que, dans le cas où vous choisissez l'option d'ouvrir directement un éditeur de textes, cet éditeur soit lancé avec des options par défaut incompatibles avec cette norme. Ceci se traduirait par un affichage erroné des caractères accentués. Si tel est le cas, nous vous conseillons de choisir l'option consistant à enregistrer les résultats dans un fichier disque. Il est alors de **votre ressort** de lancer votre éditeur de textes avec des options compatibles avec ISO 8859-1.

Aide : Affichage des fins de lignes :

Si vous choisissez l'affichage des fins de lignes, les contextes restitués respecteront les fins de lignes telles qu'elles figurent dans les textes. Dans le cas contraire, vous obtiendrez un affichage au kilomètre plus compact mais qui peut se révéler gênant pour les textes en vers.

Aide : Mise en évidence des mots cherchés

Vous avez peut-être envie que les mots spécifiés dans les séquences que vous avez cherchées soient mis en évidence dans les contextes que vous allez récupérer. Dans ce cas choisissez une des possibilités offertes.
Par exemple, si vous choisissez une mise en évidence avec **[* et *]**, et si vous aviez cherché la séquence **verre de vin** ,vous obtiendrez des contextes ayant l'allure suivante :

Et il se hâta de payer son [*verre*] [*de*] [*vin*], et il courut
au chemin de fer, voir s' il pourrait encore prendre
le train de sept heures dix.

La taille des contextes :

Suivant le choix que vous faites, vous obtiendrez des contextes plus ou moins étendus :

Taille restreinte :

Pour des textes du domaine public : Vous obtiendrez en principe le plus petit contexte (arrondi à des frontières de phrases) contenant tous les mots recherchés. Cependant, si le texte contient des phrases trop longues, ou s'il ne contient aucune ponctuation, son début et sa fin ne correspondront pas forcément à des limites de phrases.

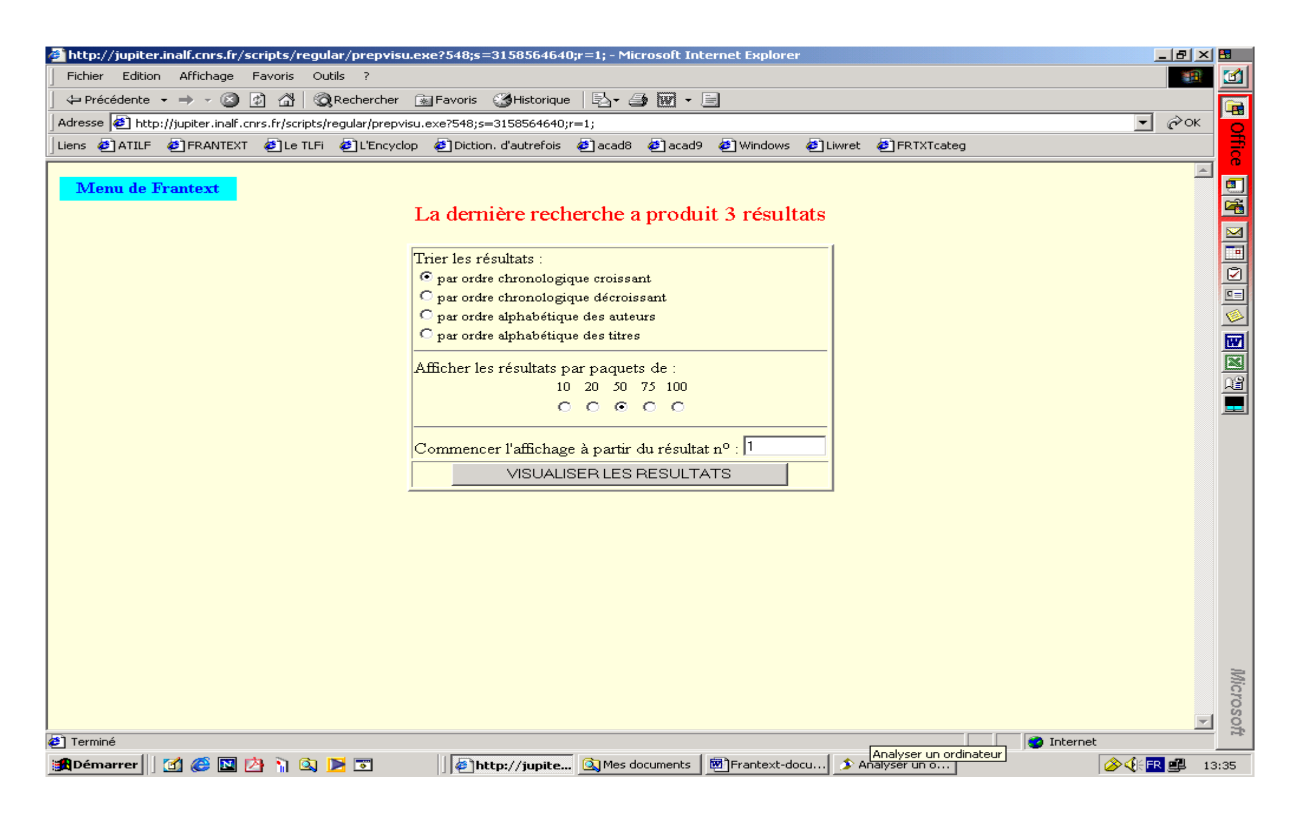
Pour les textes sous droits : Vous obtenez comme ci-dessus un contexte arrondi à des frontières de phrases. Cependant, les contextes restitués étant limités à 300 caractères (cette limite est imposée par le souci de respecter le principe des « courtes citations »), les débuts et fins de phrases peuvent être éventuellement tronqués.

Taille élargie :

Pour les textes du domaine public : En plus du contexte minimal (voir ci-dessus taille restreinte pour les textes du domaine public), le programme s'efforce de restituer la phrase précédente et la phrase suivante. Le contexte total est cependant limité à 300 mots, ce qui peut provoquer des troncatures occasionnelles.

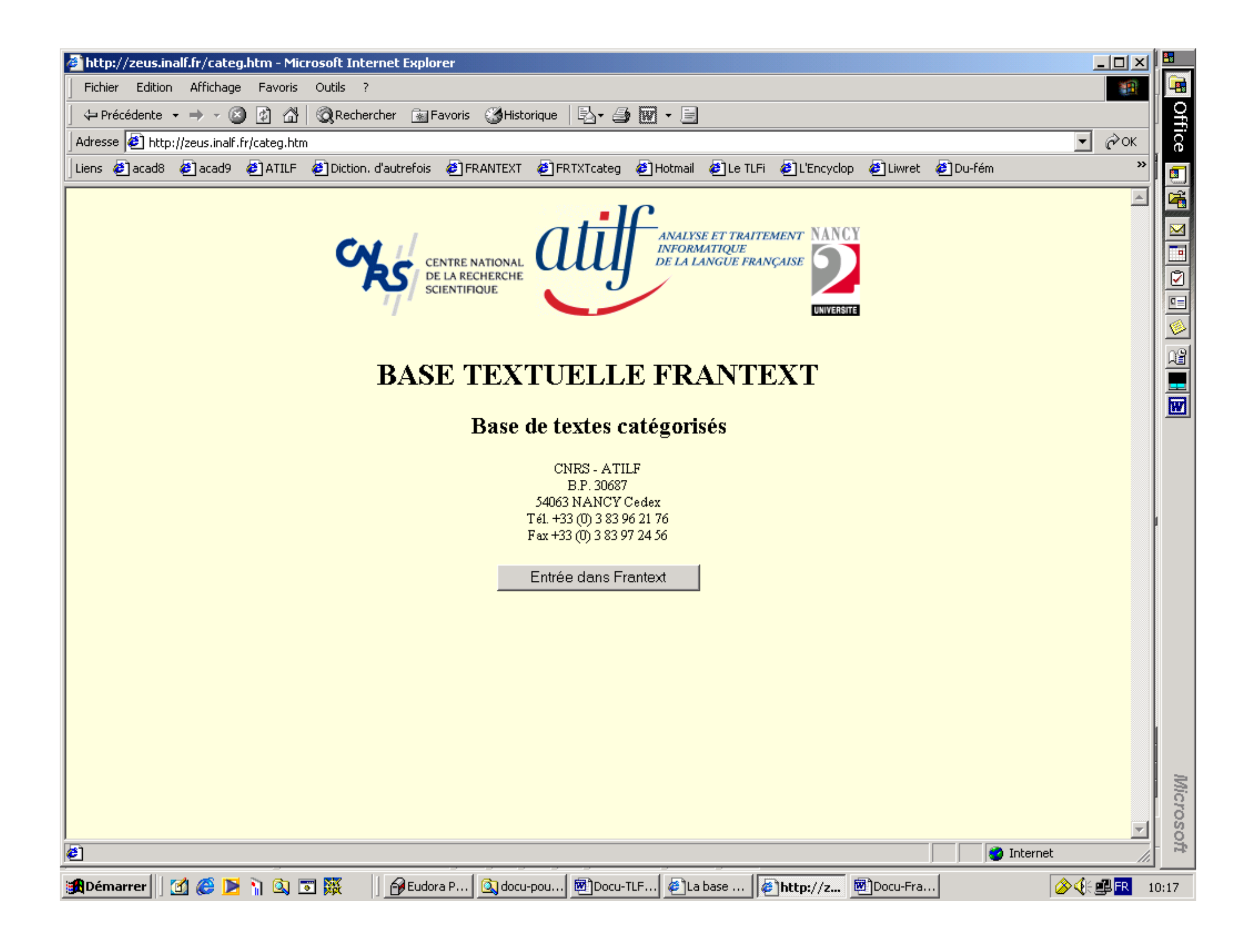
Pour les textes sous droits : Le contexte restitué est systématiquement porté à la limite maximale des 300 caractères. Le contexte restitué sera centré sur les mots recherchés. Ceci vous assure une visibilité maximale aussi bien en amont qu'en aval des mots cherchés, mais il est à noter que, fatalement, le contexte ne sera pas arrondi à des frontières de phrases.

3. Le tri des résultats

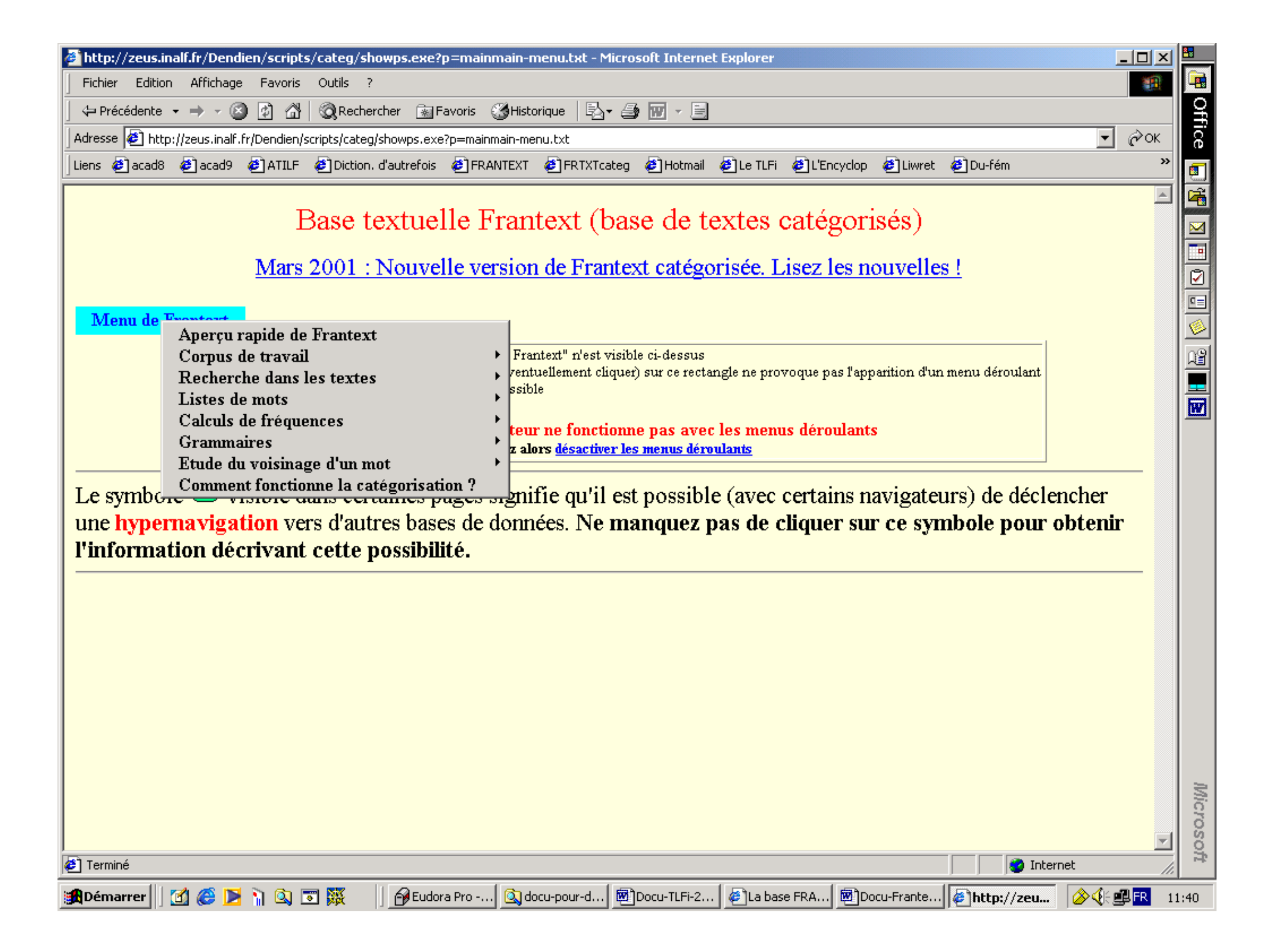


2ème section : Frantext catégorisé

Écran d’accueil sur le site FRANTEXT catégorisé



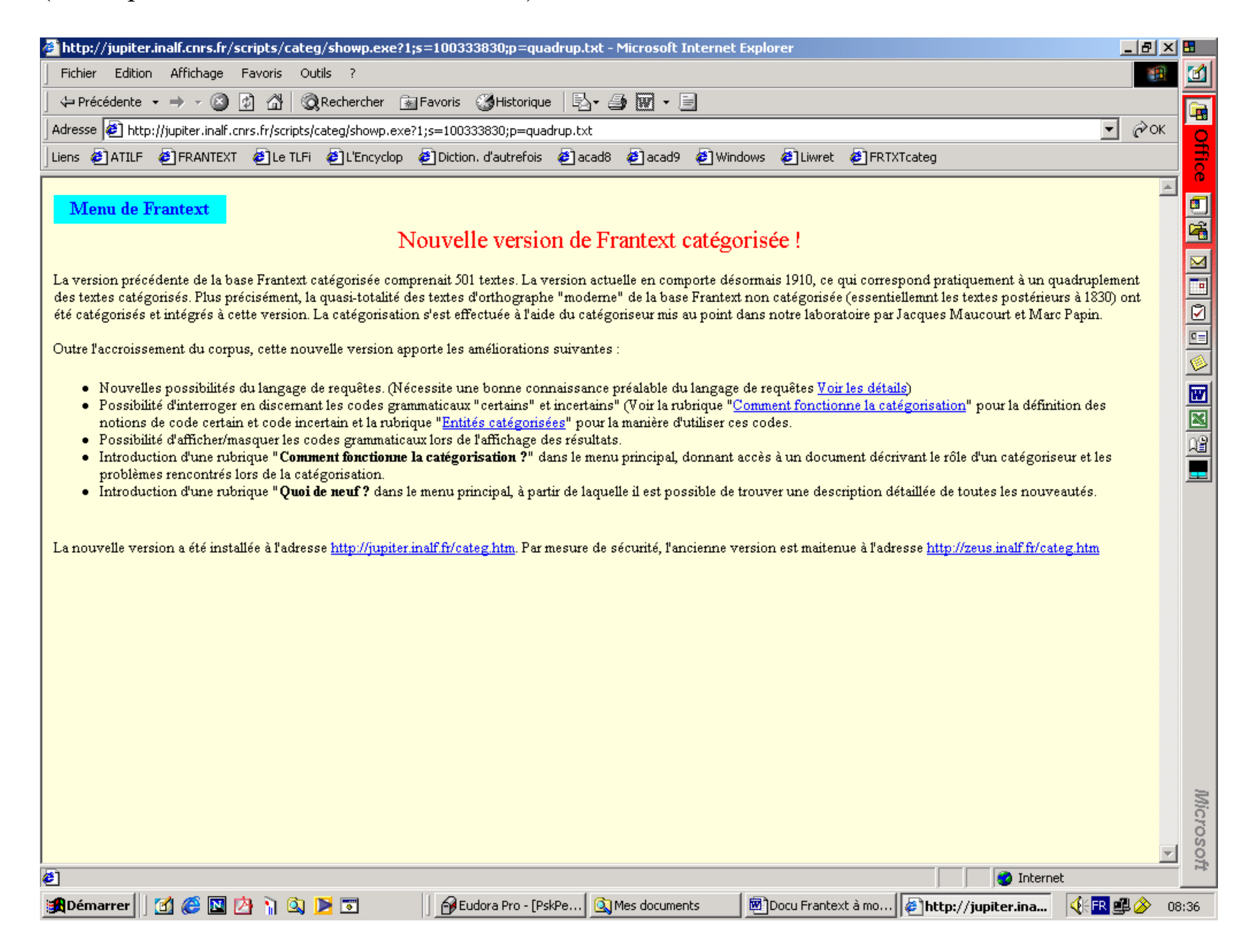
Entrée dans FRANTEXT catégorisé



Le menu proposé est globalement le même que pour Frantext non catégorisé, à la différence près qu'il s'y ajoute un bouton « Comment fonctionne la catégorisation ».

Présentation rapide de FRANTEXT catégorisé

(en cliquant sur « Lisez les nouvelles »)



« Voir les détails » renvoie à un complément concernant les expressions de séquence. Nous récapitulerons celles qui sont accessibles dans un chapitre ci-dessous.

« Comment fonctionne la catégorisation » sera aussi examiné dans un chapitre particulier.

« Entités catégorisées » et « Grammaires » font aussi l'objet de chapitres particuliers.

Pour le reste, le maniement de Frantext catégorisé est globalement le même que celui de la base non catégorisée.

Comment fonctionne la catégorisation

Rôle et limites du catégoriseur

- **Rôle du catégoriseur**
  Un catégoriseur est un programme ayant pour fonction de découper un texte en une suite de segments auxquels il va attribuer une catégorie grammaticale. Généralement un segment comprend un seul mot, mais il peut arriver qu'un segment en contienne plusieurs, dans la mesure où ce segment peut être considéré comme une entité grammaticale insécable (par exemple une locution adverbiale). Aucune norme ne définit l'ensemble des catégories attribuées : chaque catégoriseur possède donc son propre ensemble de catégories (voir l'ensemble des catégories définies dans le cas de Frantext).

  Illustrons le rôle du catégoriseur par un exemple
  Considérons la phrase "Pierre et Paul arrivèrent en même temps."
  Le découpage de ce texte par le catégoriseur, avec les codes grammaticaux associés à chaque segment sera de la forme :
  [Pierre Np] [et Cc][Paul Np] [arrivèrent V] [en même temps Adv].
  La signification des catégories est la suivante : Np=Nom propre, Cc=Conjonction de coordination, V=Verbe, Adv=Adverbe
  Les limites de segments sont matérialisées par [et].

  Les codes attribués par le catégoriseur de l'ATILF font partie intégrante de la base Frantext. Ceci rend possible la recherche de contextes contenant un séquence constituée de mots ou de catégories données. Par exemple, il est possible de rechercher les contextes contenant deux ***noms propres*** séparés par le mot ***et***, ou les contextes contenant ***une forme du verbe arriver*** suivie d'un ***adverbe***.

- **Limites du catégoriseur**
  Outre le découpage du textes en segments, le catégoriseur doit remplir un rôle de désambiguïsation des mots homographes. Par exemple le mot "entre" est parfois préposition, parfois forme du verbe "entrer", et le catégoriseur se doit de distinguer ces deux cas.
  Afin de mener à bien cette désambiguïsation, le catégoriseur applique un certain nombre de règles de décision. On comprendra sans peine que, compte tenu de l'extrême souplesse syntaxique de la langue, le catégoriseur peut se retrouver dans des situations imprévues. Elles débouchent sur deux types de problèmes :
  - Plusieurs règles de désambiguïsation du catégoriseur sont en conflit : le catégoriseur choisit alors d'appliquer la règle la plus probable : un code grammatical est attribué, mais il est marqué comme **"incertain"**.
  - Une règle de désambiguïsation du catégoriseur jugée "sûre" peut s'avérer fausse dans un contexte particulier : le code grammatical attribué soit marqué comme **"certain"**, bien qu'il soit erroné.

La nouvelle version de Frantext permet d'interroger les catégories grammaticales avec les nuances "certain" ou "incertain" : il est essentiel de bien noter la signification de ces nuances. Pour plus de détails, voir plus loin « Entités catégorisées et Grammaires ».

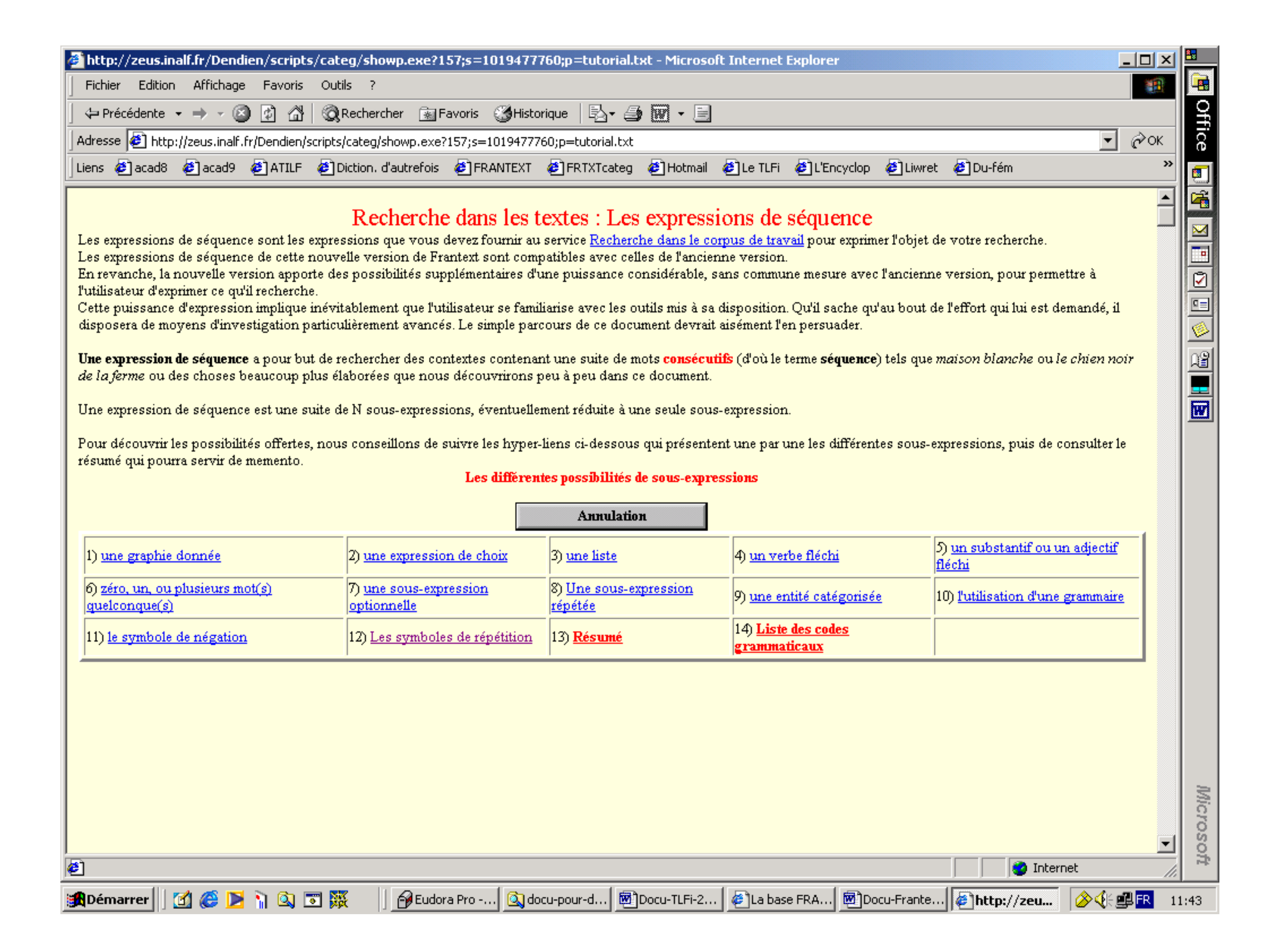
Qu'est-ce qu'un "bon" catégoriseur ?

On serait tenté de dire que la qualité d'un catégoriseur se mesure à son taux de réussite. Ceci n'est qu'en partie vrai, comme nous allons le montrer.
Prenons, par exemple le cas de nombreux mots pouvant être employés indifféremment comme adjectifs ou substantifs. Plus précisément, prenons le cas d'un mot généralement employé comme substantif et, à titre exceptionnel (disons une fois sur mille) comme adjectif. Supposons que ce mot soit employé 10000 fois dans un corpus, donc 10 fois comme adjectif. Il est alors possible de "tricher" en faisant croire au catégoriseur que le mot est un pur substantif : un tel catégoriseur se trompera en tout et pour tout 10 fois (les 10 occurrences d'adjectif qui seront à tort classées substantif).
Un revanche, un catégoriseur "honnête" aura à procéder loyalement à désambiguïsation adjectif/substantif des 10000 occurrences. On sait que la désambiguïsation adjectif/substantif est des plus délicates (que pensez vous de la phrase "***Ces puissants soporifiques nous endorment.***" ?). Supposons qu'elle échoue dans 1% des cas : le nombre d'erreurs commises est alors de 100. Pire encore, sur les 9990 cas où le mot est substantif, le catégoriseur va se tromper environ 99 fois en classant le mot comme adjectif. En définitive, le nombre de fois où le mot a été classé adjectif sera de 99 (cas d'erreurs) plus les cas (9 maximum) où la classification adjectif est correcte. On aura donc **un taux d'erreur de plus de 90% !!!**

Le catégoriseur qui se trompe dix fois moins que l'autre est-il le meilleur ? On ne peut que répondre par la négative, car ce catégoriseur masque la vérité : l'utilisateur qui désire trouver des emplois "adjectif" du mot (et donc qui désire rechercher un fait de langue précieux par sa rareté) **sera bredouille**, alors que catégoriseur honnête lui proposera cent attestations d'emploi adjectif, fausses il est vrai à environ 90%, mais **qui contiendront** les précieuses attestations cherchées. Paradoxalement, devant ces résultats, l'utilisateur non averti aura l'impression d'un catégoriseur épouvantablement mauvais (90 % d'erreurs !!!!!).

La catégorisation grammaticale est donc, comme on le voit, un art particulièrement difficile et ingrat. Nous avons néanmoins adopté le parti-pris de la transparence totale en permettant à l'utilisateur de rechercher aussi bien les codes "certains" que les codes "incertains" et en lui offrant la possibilité de rendre visibles tous les codes grammaticaux lors de l'affichage des contextes.
Il aura quelquefois l'impression que certaines erreurs auraient pu être évitées : qu'il comprenne qu'elles sont parfois engendrées par des effets secondaires de phénomènes très complexes.

Les expressions de séquence



1. Une graphie donnée

C'est la manière la plus "banale" de désigner un élément de séquence.

Exemple d'expression de séquence composée de deux graphies données:
maison blanche
Cette expression recherche des occurrences de ***maison blanche***.

2. Expression de choix

Une expression de choix s'exprime sous la forme : (Choix_1 | Choix_2 | ... | Choix_n)
dans laquelle Choix_1, Choix_1, ..., Choix_n sont des expressions de séquence.
Exemple 1 d'utilisation d'expression de choix :
maison (blanche | bleue)

Remarques :

- L'expression ci-dessus est équivalente à **(maison blanche | maison bleue)**. On notera cependant que la mise en facteur du mot "maison" a le double avantage de donner une

expression plus compacte et de procurer un temps de recherche plus bref lors de l'exécution de la requête.
- Des caractères blancs ont été ajoutés dans les exemples ci-dessus pour les rendre plus lisibles. On aurait pu les supprimer et écrire, par exemple, **maison(blanche|bleue)**

Exemple 2 :
Puisqu'une expression de choix est une expression de séquence et qu'un des choix est lui-même une expression de séquence, on en déduit que l'on peut imbriquer les expressions de choix à volonté.
(maison|palais)d'un(blanc(immaculé|sale)|bleu(d'azur|profond))
est une expression qui cherchera les occurrences de ***maison d'un blanc immaculé***, ***palais d'un blanc immaculé***, ***maison d'un blanc sale***, ***palais d'un blanc sale***, ***maison d'un bleu d'azur***, ***palais d'un bleu d'azur***, ***maison d'un bleu profond***, ***palais d'un bleu profond***

3. Les listes

&lxxx désigne un des mots de la liste xxx.
(On se reportera au service création/édition de listes de mots pour obtenir une description des différentes manières de créer une liste de mots.)
Exemple :
Si dans une liste de nom *couleur* on a les mots rouge, vert, jaune, on peut écrire l’expression de séquence suivante :
volet &lcouleur
pour chercher *volet rouge, volet vert, volet jaune*

Remarque : une liste est une suite d’éléments constitués d’une seule graphie (le « blanc » est considéré comme un séparateur de mots). Un élément de liste ne peut donc pas être par exemple « vert bouteille ». Cette restriction est compensée par le recours aux grammaires présentées plus loin dans ce document.

4. Verbe fléchi

&cxxx
désigne une des formes fléchies du verbe dont l'infinitif est xxx
Exemple d'expression de séquence :
&caimer les bons repas

5. Substantif ou adjectif fléchi

&mxxx
désigne une des formes d'un substantif dont le singulier est xxx, d'un adjectif dont le masculin singulier est xxx.
Exemple d'expression de séquence :
&mfleur &mvert
Cette séquence trouvera les occurrences de ***fleur verte*** ou ***fleurs vertes***

6. Zéro, un ou plusieurs mot(s) quelconque(s)

&q(n_1,n_2)
désigne une suite de mots dont le nombre est compris entre n_1 et n_2 .
Exemple d'expression de séquence :
un &q(0,2) homme
Cette expression de séquence pourra trouver des contextes tels que ***un*** soit séparé de ***homme*** par zéro, un ou deux mots, donc des contextes tels que ***un homme***, ***un grand homme***, ***un très petithomme***.
Notes :

- Les nombres n_1 et n_2 doivent respecter les règles suivantes :
  - n_1 doit être positif ou nul. n_2 doit être strictement positif.
  - n_2 doit être supérieur à n_1
  - La différence n_2-n_1 doit être inférieure ou égale à 7 (ce qui permet une amplitude de 8 mots). Cette restriction est due au souci de maintenir un temps de recherche raisonnable en évitant au logiciel d'avoir à faire des hypothèses trop nombreuses. Il est demandé de ne pas essayer de contourner cette restriction avec une expression du genre **un&q(0,7) &q(0,7)homme** qui conduirait effectivement à tolérer jusqu'à 16 mots entre ***un*** et ***homme***, mais au prix d'une violente dégradation du temps de réponse dont l'utilisateur serait la première victime. La bonne solution pour chercher des séquences distantes de plus de 8 mots est de chercher deux séquences en cooccurrence (voir aide sur cette rubrique, cadre 2 du formulaire de recherche).
- L'expression **&q** utilisable dans les versions antérieures et qui désigne un mot quelconque est toujours utilisable. Elle est strictement équivalente à **&q(1,1)**.

7. Sous expressions optionnelles

Supposons que l'on veuille chercher des contextes contenant ***un homme*** ou ***un grand homme***.
Une telle recherche peut s'exprimer avec **(un | un grand) homme** ou avec **un (homme | grand homme)** en utilisant une expression de choix.
De telles expressions sont lourdes et inélégantes.

Nous proposons donc de les simplifier avec le symbole **&?** dont la signification est la suivante :
Si **&?** est situé devant une expression de séquence, alors cette expression est **facultative** dans les contextes recherchés.

Exemples :

- **un &?grand homme** recherchera les contextes ***un homme*** ou ***un grand homme***.
- **un &?(très grand) homme** recherchera les contextes ***un homme*** ou ***un très grand homme***.
- **un &?(&?très grand) homme** recherchera les contextes ***un homme*** ou ***un grand homme*** ou ***un très grand homme***.
-

8. Les sous-expressions répétées

(Points 8 et 12 du tableau en ligne)
Supposons que l'on recherche une énumération d'adjectifs séparés par des virgules. Une telle énumération est la répétition d'une sous-expression de la forme "**, &e(g=A)**" (dans laquelle **&e(g=A)** désigne un adjectif). Le problème est que le nombre de répétitions étant inconnu, il n'est pas possible de l'exprimer sans introduire un symbole spécial dans le langage de requêtes. Ce symbole est noté **&+**.
Placé devant un élément simple (comme dans l'expression **&+ nous**), il signifie que cet élément peut se répéter une ou plusieurs fois. Placé devant un bloc parenthésé, il signifie que c'est tout le bloc qui peut se répéter une ou plusieurs fois. Par exemple **&+(, &e(g=A))** signifie que c'est la séquence "virgule - adjectif" qui se répète.
Une expression décrivant une répétition d'au moins deux adjectifs s'écrit donc : **, &e(g=A) &e(g=A) &+(, &e(g=A))**

Un autre symbole de répétition, noté **&*** existe également. Il a la même signification que **&+**, à l'exception du fait que la répétition peut se faire zéro, une ou plusieurs fois. Ainsi **&e(g=S) &* &e(g=A)** désigne un substantif suivi éventuellement d'un ou plusieurs adjectifs.

Les symboles de répétition peuvent se trouver devant une négation. Ceci n'est pas sans poser problème. Considérons en effet l'expression **&+ ^amour**. Elle décrit une séquence de mots **dont aucun n'est le mot amour**. Son utilisation seule (bien qu'elle ne soit pas interdite) n'a donc guère de sens. De même son utilisation en début d'expression (par ex. **&+ ^amour fou**) ou en fin d'expression (par ex. **mon &+^amour**) sont à utiliser avec circonspection, car les contextes correspondant à de telles expressions ne sont bornés, soit à gauche, soit à droite que par la prochaine occurrence de "amour" qui peut être très lointaine.
Même lorsqu'elle est utilisée au milieu d'une expression, cette possibilité doit être utilisée avec prudence, en raison d'un phénomène que nous allons décrire. Par exemple, supposons que l'on veuille rechercher les contextes contenant tous les passages entre parenthèses.
On serait tenté d'écrire l'expression : **\(&+^\) \)** (on rappelle que dans le langage de requêtes les parenthèses sont réservées aux expressions de listes et de blocs parenthésés et que, donc, si l'on veut rechercher une parenthèse en tant que texte, il faut la faire précéder du caractère "\"). Cependant, cette expression présente un défaut, car elle part du préjugé que le texte est lu de la gauche vers la droite. Or le logiciel Stella applique des algorithmes complexes d'optimisation qui l'amènent à lire indifféremment dans les deux sens. Si, lors de la recherche de l'expression ci-dessus, la lecture se fait de la gauche vers la droite, tout se passera bien. En revanche, si la lecture se fait de la droite vers la gauche, le logiciel partira de la parenthèse fermante, puis se propagera vers la gauche **tant qu'il ne rencontrera pas de parenthèse fermante**. Il passera donc "au-dessus" de la parenthèse ouvrante.
Pour empêcher se problème de survenir, il convient donc de veiller à éviter une propagation incontrôlée dans un sens comma dans l'autre. La solution consiste à écrire l'expression sous la forme **\(&+^(\(| \)) \)** qui arrête la propagation dès qu'une parenthèse ouvrante ou fermante est rencontrée. Dans le cas où le texte contiendrait une parenthèse ouvrante/fermante sans fermante/ouvrante correspondante dans son voisinage, Stella arrête automatiquement la propagation au bout de cent répétitions.

Terminons ce chapitre par le problème suivant : trouver toutes les phrases qui contiennent une séquence du genre ***en , en ..., en .... etc.*** (... désigne une suite de mots quelconques à l'exclusion d'une virgule. Par ex. ***en flânant, en ne se pressant pas, en s'amusant etc.***).
La sous-expression qui se répète peut s'écrire **, en &+^,** . Cependant, cette sous expression n'est bornée à droite que par la prochaine apparition d'une virgule. Elle peut donc outrepasser

le point final de phrase, ce que nous ne voulons pas. Afin d'éviter ceci, il suffit de réécrire la sous expression sous la forme **, en &+^(,|.)**. En ajoutant le symbole de répétition, ceci devient **&+(, en &+^(,|.))** . En rajoutant devant cette expression le "en ..." initial, ou aboutit à la requête : en **&+^(,|.) &+(, en &+^(,|.))**

9. Entités catégorisées

(voir chapitre suivant, §1)

10. Grammaires

(voir chapitre suivant,§2)

11. Le symbole de négation

(Point 11 du tableau en ligne)
Le symbole de négation est le caractère ^. Il peut être placé devant n'importe quel type de sous expression. Quel que soit le type de sous-expression, l'ensemble formé par le symbole de négation et la sous-expression désigne **une graphie** qui est tout sauf le point de départ de la sous-expression.

Exemples :

- **homme ^très grand qui** est une expression de séquence qui va chercher tous les contextes du genre ***homme XXX grand qui*** tels que XXX ne soit pas égal à ***très***.
- **homme ^(très grand)** est une expression de séquence qui va chercher tous les contextes du genre ***homme XXX*** tels que ou bien XXX est différent de ***très***, ou bien XXX est égal à ***très*** et le mot suivant est différent de ***grand*** (autrement dit XXX n'est pas le point de départ d'une séquence ***très grand***).
- **&carriver ^&e(g=A) à** est une expression de séquence qui va sélectionner des contextes tels que ***arriveront vite à***, mais pas un contexte tel que ***arrivaient en même temps à***. En effet **^&e(g=A)** désigne une **graphie** qui n'est pas un adjectif, mais ne désigne pas **une entité catégorisée** qui n'est pas un adjectif : cette dernière désignation se noterait en effet &e(g!=A). On prendra donc bien garde à ne pas confondre ^&e(g=A) et &e(g!=A).

12. Les symboles de répétition

(voir les sous expressions répétées)

13. Résumé

(voir chapitre suivant, §3)

14. Liste des codes

(voir chapitre suivant, §1.3)

Entités catégorisées et Grammaires

1.Entités catégorisées :

1.1. Préambule :

Attention : Cette possibilité n'est offerte que sur les versions de Frantext pour lesquels la mention "Base de textes catégorisés" apparaît dans le menu principal.
Un corpus catégorisé est un corpus qui a été traité par un programme de catégorisation (appelé **catégoriseur**) avant sa mise sous forme de base textuelle.
Le rôle du catégoriseur est de découper les textes en une suite d'entités auxquelles il attribue un code grammatical (verbe, adverbe, adjectif,...). Si, dans la majorité des cas, il y a coïncidence entre graphie et entité, il n'en est pas toujours ainsi. Par exemple, le mot composé ***tire-bouchon*** est considéré par le catégoriseur comme une seule entité ayant le code substantif, alors qu'il est composé de trois graphies. De même, certains groupes de graphies comme ***en même temps*** sont considérés comme une seule entité (code adverbe).
Afin de manipuler ces entités catégorisées, une nouvelle expression de séquence est introduite :

&e(xxx)
désigne une entité catégorisée. La chaîne de caractères xxx désigne les **propriétés** de l'entité catégorisée recherchée.

1.2. Les différentes possibilités :

- **&e()**
  Cette expression désigne une entité catégorisée quelconque. Elle constitue donc un véritable "joker" d'entité catégorisée.
  Exemple :
  &carriver &e() à
  est une expression qui trouveront des contextes tels que ***arriveront jamais à***, mais aussi ***arrivaient en même temps à***
- **&e(g=*yyy*)**
  désigne une entité ayant le code grammatical précisé par *yyy*, où *yyy* est une chaîne de caractères formée par un ou plusieurs codes grammaticaux.
  Exemples :
  - **&e(g=S)** désigne un substantif.
  - **&e(g=S A)** désigne un substantif **ou** un adjectif.

  On se reportera à la liste des codes disponibles pour connaître l'ensemble des codes existants.

  Codes « certains » et « incertains » : **Tout code grammatical** peut être employé tel quel ou peut être préfixé par la lettre "i" ou "c".
  Ainsi

  - **&e(g=cS)** désigne les substantifs qui ont été reconnus "avec certitude" par le programme de catégorisation.
  - **&e(g=iS)** désigne les substantifs qui ont été reconnus "avec incertitude" par le programme de catégorisation.
  - **&e(g=S)** désigne les substantifs qui ont été reconnus "avec certitude ou avec incertitude" par le programme de catégorisation.

- **&e(g!=*yyy*)**
  désigne une entité **n'ayant pas** le code grammatical précisé par *yyy*.
  Exemples :
  - **&e(g!=S)** désigne une entité autre que substantif.
  - **&e(g!=S A)** désigne une entité autre que substantif **ou** adjectif.

  (Même remarque que ci-dessus pour les codes « certains » ou « incertains »)

- **&e(c=*expression*)**
  désigne une entité catégorisée ayant un contenu textuel conforme à ***expression***. Une expression de contenu textuel est une expression de séquence qui ne comporte aucune expression **&e(xxx)** : en effet, les graphies contenues dans une entité catégorisée ne sont pas elles-mêmes catégorisées.
  Exemples :
  - **&e(c=tire-)** recherche les entités catégorisées contenant ***tire-***
  - **&e(c=(&mdemi &q|&q &mdemi))** recherche les entités catégorisées contenant au moins deux mots, dont ***demi***, ***demis***, ***demie*** ou ***demies***
- **&e(c!=*expression*)** est une entité **n'ayant pas** le contenu indiqué par ***expression***.
- Expressions d'entités composites.
  Toutes les possibilités ci-dessus peuvent se combiner entre elles.
  Exemple :
  - **&e(c=(&mdemi &q | &q &mdemi) c!=&mheure g=S)** désigne une entité comportant au moins deux mots (dont ***demi*** fléchi), mais ne contenant ni ***heure***, ni ***heures***, et qui de surcroît soit un substantif.
    On notera que l'ordre des expressions **c=...**, **g=...** au sein d'une même entité catégorisée est indifférent. On aurait pu écrire l'entité ci-dessus sous la forme strictement équivalente :
    &e(g=S c!=&mheure c=(&mdemi &q | &q &mdemi))

1.3. Liste des codes grammaticaux (établie provisoirement)

Attention : Ces codes sont à utiliser "tels quels" en respectant majuscules et minuscules. Ils peuvent être éventuellement précédés des préfixes "i" ou "c"
Par ex.

- **&e(g=cCc)** désigne une conjonction de coordination reconnue "avec certitude" par le programme de catégorisation.
- **&e(g=iCc)** désigne une conjonction de coordination reconnue "avec incertitude" par le programme de catégorisation.
- **&e(g=Cc)** désigne une conjonction de coordination reconnue "avec ou sans certitude" par le programme de catégorisation.

A	adjectif (sauf cas Aca, Apr, Aps)
Aca	adjectif cardinal
APr	adjectif/participe présent
APs	adjectif/part. passé
Adv	Adverbe
Avn	Partie d'une négation (par ex. cas de **ne / n'**, ou **pas / point / guère ...** si associés à **ne** ou **n'**)
Cc	Conjonction coordination
Cs	Conjonction subordination
D	Déterminant (sauf cas Dca, Dg)
Dca	car. dét (cardinal ayant le rôle d'un déterminant : *deux* pigeons s'aimaient)
Dg	amalgamés (au/aux/du/des)
E	exclamatif
Ep	présentatif (voici, voilà, ...)
Ger	gérondif (*en* lié à un participe présent)
Inf	infinitif
Inj	interjection (ah, oh, ha, ho, ...)
Int	interrogatif
Np	Nom propre
Nu	numeral card.
Ono	onomatopée
P	Pronom (sauf cas Per, X)
Per	Pronom personnel
Pp	Préposition
Pr	Participe présent sauf cas APr, Ger
Ps	Participe passé (sauf cas APs)
S	Substantif
V	Verbe (sauf participes et infinitif)
R	mot inconnu du logiciel
X	mot non traité (que/qu', où, sinon)

Remarque importante : il n'existe à ce jour aucun document décrivant les critères linguistiques qui sous-tendent cette catégorisation (utilisée dans le dictionnaire TLFnome de l'ATILF), ni les critères d'attribution de ces étiquettes.

2. Les grammaires

2.1. Préambule

Les grammaires, telles que nous allons les définir, permettent de formuler de puissantes expressions de recherche capables de localiser dans un corpus des occurrences de phénomènes multiformes : par exemple, une référence à une période peut prendre la forme ***dès 1954***, ou ***en juillet dernier/prochain***, ***la semaine dernière/prochaine***, etc.
Les grammaires (au sens où nous l'entendons, qui est celui de la théorie des langages) ne sont pas à confondre avec les grammaires des langues naturelles. En revanche, elles permettent de rechercher des phénomènes (par exemple une construction pronominale de verbe) qui peuvent se manifester sous des formes très variables en raison des règles syntaxiques du français. En ce sens, il y a bien un lien conceptuel entre les grammaires que nous proposons et la grammaire française.
Malgré ce lien, les linguistes et les grammairiens comprendront que :

- nous donnons aux termes **grammaire** et **règle de grammaire** une signification différente de celle à laquelle ils sont habitués, mais qui est conforme à la terminologie de la théorie des langages.
- nous ne prétendons pas résoudre les difficultés des langues naturelles en proposant un simple outil : ce n'est pas parce que nous offrons une paire de ciseaux (nos grammaires) que nous prétendons être tailleurs (grammairiens)

2.2.Qu'est-ce qu'une grammaire ?

Une grammaire est un fichier qui contient des **règles**. Ce fichier peut être saisi par l'utilisateur grâce au service de création de grammaire.

L'intérêt des grammaires est triple :

- Une grammaire est un recueil d'expression de séquences tapées une fois pour toutes par l'utilisateur.
- Les expressions de séquences peuvent être paramétrées, ce qui permet de les utiliser pour des recherches multiples.
- Une grammaire permet d'élaborer facilement des expressions de séquences d'une complexité arbitrairement élevée.

Chaque règle de grammaire a un **nom** et un **corps**.

2.3. Un premier exemple de grammaire

Voici un exemple de grammaire : (**En gras : nom de la règle**, en italique : *corps de la règle*)

quantifieur :
très | assez | excessivement
qualifieur :
grand | gros | petit
qualification :
&rqualifieur | &rquantifieur &rqualifieur

Supposons que vous ayez saisi une telle grammaire dans un fichier de nom **xxx**. Alors, vous pourrez utiliser les règles de cette grammaire dans des expressions de recherche lorsque vous ferez appel au service recherche dans le corpus de travail. Pour utiliser une règle de la

grammaire xxx (règle "quantifieur" par exemple), vous taperez ***&rquantifieur,xxx***. Une telle expression est appelée **invocation** de la règle "quantifieur" de la grammaire xxx.
Elle est équivalente à l'expression obtenue en plaçant le corps de la règle entre parenthèses :

&rquantifieur,xxx<=>***(très | assez | excessivement)***

Vous pourrez par exemple, avec la séquence ***&rquantifieur,xxx &rqualifieur,xxx homme*** chercher des contextes contenant *très grand homme*, *assez petit homme*, etc...

De même, en utilisant la règle **qualification** vous pourrez, avec la séquence ***&rqualification,xxx homme*** chercher des contextes contenant *très grand homme* (ou simplement *grand homme*), *assez petit homme* (ou simplement *petit homme*), etc...
Il est à noter que la règle **qualification** invoque les règles **quantifieur** et **qualifieur** sans spécifier de nom de grammaire : en effet le nom de la grammaire est facultatif à l'intérieur d'une même grammaire.

2.4. Les grammaires, c'est facile à écrire !

Voyons, sur un exemple, que l'écriture d'une grammaire est chose très facile. Proposons-nous d'écrire une grammaire capable de rechercher les contextes contenant des **constructions pronominales** de verbes. Par simplification, nous laisserons de côté les phrases interrogatives et supposerons que le sujet du verbe est un pronom.

Les usages pronominaux sont :

- temps simple : ***Je me/m'*** (ou ***tu te/t'***, ***il/elle/ils/elles/on/qui se/s'***, ***nous nous***, ***vous vous***) + **une forme verbale**
- temps composé : ***Je me/m'*** (ou ***tu te/t'***, ***il/elle/ils/elles/on/qui se/s'***, ***nous nous***, ***vous vous***) + **auxiliaire être** + **participe passé**
- temps simple avec négation : ***Je ne me/m'*** (ou ***tu ne te/t'***, ***il/elle/ils/elles/on/qui ne se/s'***, ***nous ne nous***, ***vous ne vous***) + **une forme verbale** + **pas/plus/jamais/guère/point/mie/etc.**
- temps composé avec négation : ***Je ne me/m'*** (ou ***tu ne te/t'***, ***il/elle/ils/elles/on/qui ne se/s'***, ***nous ne nous***, ***vous ne vous***) + **auxiliaire être** + **pas/plus/jamais/guère/point/mie/etc.** + **participe passé**

On constate que temps simple et temps composé (sans négation) commencent de la même façon, que nous pouvons décrire avec une première règle de grammaire de nom **affirmatif** :

affirmatif :
je (me|m') | tu(te|t') | (il|elle|ils|elles|on|qui) (se|s') |
nous nous | vous vous

Remarque : le corps d'une règle peut s'étendre sur plusieurs lignes (un retour à la ligne n'a aucune signification).
Il nous est maintenant facile d'écrire la règle **temps_simple_affirmatif** pour les temps simples :
temps_simple_affirmatif :
&raffirmatif &e(g=V)
De même nous pouvons écrire la règle **temps_compose_affirmatif** pour les temps composés :

temps_compose_affirmatif :
&raffirmatif &cêtre &e(g=Ps)
De la même façon, nous écrivons la suite de la grammaire :
negatif :
je ne (me|m') | tu ne (te | t') | (il|elle|ils|elles|on|qui) ne (se|s') |
nous ne nous | vous ne vous
temps_simple_negatif :
&rnegatif &e(g=V) &rfin_negation
Remarque : nous préciserons plus loin que la règle **fin_negation** correspond à pas/plus/jamais/guère/point/mie/etc.
temps_compose_negatif :
&rnegatif &cêtre &rfin_negation &e(g=Ps)
fin_negation :
pas|plus|jamais|guère|mie|point
Enfin, il nous reste à réunir les quatre cas de figure en une seule règle :
usage_pronominal :
&rtemps_simple_affirmatif | &rtemps_compose_affirmatif |
&rtemps_simple_negatif | &rtemps_compose_negatif

Voici donc notre grammaire construite :

```
affirmatif :
    je (me|m') | tu(te|t') | (il|elle|ils|elles|on|qui) (se|s') | nous nous | vous vous
temps_simple_affirmatif :
    &raffirmatif &e(g=V)
temps_compose_affirmatif :
    &raffirmatif &cêtre &e(g=Ps)
negatif :
    je ne (me|m') | tu ne (te|t') |(il|elle|ils|elles|on|qui) ne (se|s') | nous ne nous | vous ne vous
temps_simple_negatif :
    &rnegatif &e(g=V) &rfin_negation
temps_compose_negatif :
    &rnegatif &cêtre &rfin_negation &e(g=Ps)
fin_negation :
    pas|plus|jamais|guère|mie|point
usage_pronominal :
    &rtemps_simple_affirmatif | &rtemps_compose_affirmatif |
    &rtemps_simple_negatif | &rtemps_compose_negatif
```

Nous avons écrit cette grammaire pour fonctionner avec tous les verbes. Supposons maintenant que nous voulions une grammaire similaire, mais qui ne s'appliquerait qu'au verbe **laver**.
A chaque expression ***&e(g=V)*** (forme verbale quelconque) , il convient de substituer ***&claver*** (forme verbale du verbe laver). De même, à chaque expression ***&e(g=Ps)*** (participe passé d'un verbe quelconque), il convient de substituer ***&e(g=Ps c=&claver)*** (participe passé du verbe laver).
Nous obtenons donc la grammaire suivante :

```
affirmatif :
    je (me|m') | tu(te|t') | (il|elle|ils|elles|on|qui) (se|s') | nous nous | vous vous
temps_simple_affirmatif :
    &raffirmatif &claver
temps_compose_affirmatif :
    &raffirmatif &cêtre &e(g=Ps c=&claver)
negatif :
    je ne (me|m') | tu ne (te|t') |(il|elle|ils|elles|on|qui) ne (se|s') | nous ne nous | vous ne vous
temps_simple_negatif :
    &rnegatif &claver &rfin_negation
temps_compose_negatif :
    &rnegatif &cêtre &rfin_negation &e(g=Ps c=&claver)
fin_negation :
    pas|plus|jamais|guère|mie|point
usage_pronominal :
    &rtemps_simple_affirmatif | &rtemps_compose_affirmatif |
    &rtemps_simple_negatif | &rtemps_compose_negatif
```

La première objection qui vient à l'esprit est la suivante :
Vais-je devoir écrire une grammaire différente pour chaque verbe que je veux étudier ?
Ce serait évidemment très fâcheux si un nouveau mécanisme n'était pas introduit : les règles avec paramètres

2.5. Les règles avec paramètres

Nous allons élargir la syntaxe des expressions d'invocation de règles avec la possibilité suivante : ***&rXXX(P_1,P_2,...,P_n),YYY***
Cette expression invoque la règle **XXX** de la grammaire **YYY**. $P_1,P_2,...,P_n$ sont des chaînes de caractères qui sont transmises à la règle **XXX**, c'est-à-dire que, dans l'écriture du corps de **XXX**, nous pourrons faire référence à ces paramètres.

Le plus facile est d'étudier un exemple très simple : supposons que nous voulions successivement chercher les usages de toute une série de verbes donnés (aimer, danser, boire, ...) à la première personne du pluriel.
Il serait possible de formuler la requête suivante :
(nous &caimer | nous n'&caimer | nous &cavoir &e(g=Ps c=&caimer) | nous n'&cavoir &q &e(g=Ps c=&caimer) | &caimer-nous)
autant de fois qu'il y a de verbes à étudier. Une autre possibilité serait d'écrire une grammaire pour le verbe **aimer**, puis de la modifier pour **danser**, etc.
L'une et l'autre de ces solutions sont également fastidieuses.
Au lieu de cela nous allons écrire la grammaire **G** suivante :

```
auxiliaire :
    &cavoir | &cêtre
verbe :
    nous &c&1 | nous (ne|n')&c&1 | nous &rauxiliaire &e(g=Ps c=&c&1) |
    nous (ne|n') &rauxiliaire &q &e(g=Ps c=&c&1) | &c&1-nous
```

On remarquera dans cette grammaire, un nouvel élément de syntaxe : **&1**.
Cette expression signifie que la règle **verbe** doit être appelée avec au moins un paramètre, et que ce paramètre se substituera à **&1**. Autrement dit, **&1** est une manière **générique** de

désigner le premier paramètre avec lequel la règle **verbe** est invoquée.
Si maintenant, nous tapons l'expression de recherche :
&rverbe(aimer),G
la règle **verbe** est équivalente à :
nous &caimer | nous (ne|n')&caimer | nous &rauxiliaire &e(g=Ps c=&caimer) |
nous (ne|n') &rauxiliaire &q &e(g=Ps c=&caimer) | &caimer-nous
et on procédera donc à la recherche des emplois de **aimer** à la première personne du pluriel.
Bien entendu, une règle peut avoir plusieurs paramètres. Il seront respectivement désignés dans le corps de la règle par **&1**,**&2**, **&3**, etc.

Une règle peut transmettre à une autre règle un ou plusieurs des paramètres qu'elle a reçus.
Pour illustrer ceci, voyons comment nous pouvons modifier la grammaire cherchant les usages pronominaux pour la faire fonctionner sur un verbe passé en paramètre.
La règle **usage_pronominal** doit recevoir nécessairement un paramètre : le verbe à étudier.
Elle devra retransmettre ce paramètre à la règle **temps_simple_affirmatif** qui en a besoin pour savoir quel verbe il doit conjuguer. De même, la règle **usage_pronominal** devra transmettre son paramètre aux autres règles qu'elle invoque.
La grammaire devient ainsi :

```
affirmatif :
    je (me|m') | tu(te|t') | (il|elle|ils|elles|on|qui) (se|s') | nous nous | vous vous
temps_simple_affirmatif :
    &raffirmatif &c&1
temps_compose_affirmatif :
    &raffirmatif &cêtre &e(g=Ps c=&c&1)
negatif :
    je ne (me|m') | tu ne (te|t') |(il|elle|ils|elles|on|qui) ne (se|s') | nous ne nous | vous ne vous
temps_simple_negatif :
    &rnegatif &c&1 &rfin_negation
temps_compose_negatif :
    &rnegatif &cêtre &rfin_negation &e(g=Ps c=&c&1)
fin_negation :
    pas|plus|jamais|guère|mie|point
usage_pronominal :
    &rtemps_simple_affirmatif(&1) | &rtemps_compose_affirmatif(&1) |
    &rtemps_simple_negatif(&1) | &rtemps_compose_negatif(&1)
```

Essayons enfin de faire encore plus fort en réécrivant la grammaire ci-dessus pour qu'elle fonctionne, au choix, ou bien sur tous les verbes ou bien sur un verbe passé en paramètre.
Pour obtenir ceci, il faut que certaines règles changent de comportement en fonction des paramètres qui lui sont passés.
Par exemple, si nous considérons la règle **temps_simple_affirmatif**, son corps devrait être :

- tantôt du genre *&raffirmatif &e(g=V)*
- tantôt du genre *&raffirmatif &cXXX* avec XXX=verbe à conjuguer.

Supposons que nous écrivions cette règle sous la forme :

```
temps_simple_affirmatif :
&raffirmatif &rverbe_&1(&2)
```

- Si maintenant, nous invoquons cette règle en lui passant deux paramètres **particulier** et **laver**, alors son corps devient équivalent à :
  &raffirmatif &rverbe_particulier(laver)
  Maintenant nous ajoutons à la grammaire la règle :
  verbe_particulier :
  &c&1
  Alors, la chaîne **laver** est transmise à la règle **verbe_particulier** et nous obtenons un fonctionnement correct pour un verbe particulier.
- Si au contraire, nous invoquons la règle **temps_simple_affirmatif** en lui passant deux paramètres **general** et **n'importe quoi**, son corps devient équivalent à :
  &raffirmatif &rverbe_general(n'importe quoi)
  Maintenant,ajoutons à la grammaire la règle :
  verbe_general :
  &e(g=V)
  Le paramètre **n'importe quoi** sera ignoré par cette règle et nous obtenons un fonctionnement correct pour le cas général.

Nous voyons ainsi qu'il est possible de passer en paramètre absolument tout ce que l'on veut, y compris des noms ou des parties de noms de règles. Ceci permet d'écrire des règles qui vont invoquer telle ou telle règle en fonction des paramètres avec lesquels elles sont invoquées, et d'obtenir ainsi des grammaires d'une très grande souplesse.
En systématisant ce raisonnement, nous obtenons finalement la grammaire suivante :

```
affirmatif :
    je (me|m') | tu(te|t') | (il|elle|ils|elles|on|qui) (se|s') | nous nous | vous vous
temps_simple_affirmatif :
    &raffirmatif &rverbe_&1(&2)
temps_compose_affirmatif :
    &raffirmatif &cêtre &rparticipe_&1(&2)
negatif :
    je ne (me|m') | tu ne (te|t') |(il|elle|ils|elles|on|qui) ne (se|s') | nous ne nous | vous ne vous
temps_simple_negatif :
    &rnegatif &rverbe_&1(&2) &rfin_negation
temps_compose_negatif :
    &rnegatif &cêtre &rfin_negation &rparticipe_&1(&2)
fin_negation :
    pas|plus|jamais|guère|mie|point
usage_pronominal :
    &rtemps_simple_affirmatif(&1,&2) | &rtemps_compose_affirmatif(&1,&2) |
    &rtemps_simple_negatif(&1,&2) | &rtemps_compose_negatif(&1,&2)
verbe_general :
    &e(g=V)
participe_general :
    &e(g=Ps)
verbe_particulier :
    &c&1
participe_particulier :
    &e(g=Ps c=&c&1)
```

Si maintenant, nous lançons une recherche avec les invocations :

- ***&rusage_pronominal(particulier,laver),nom-de-la-grammaire*** alors nous recherchons des usages pronominaux du verbe **laver**.
- ***&rusage_pronominal(general,),nom-de-la-grammaire*** alors nous recherchons des usages pronominaux de n'importe quel verbe.
  Remarque : on notera dans l'invocation ci-dessus que le second paramètre est une chaîne vide (ce paramètre est en effet ignoré si le premier paramètre est **general**).

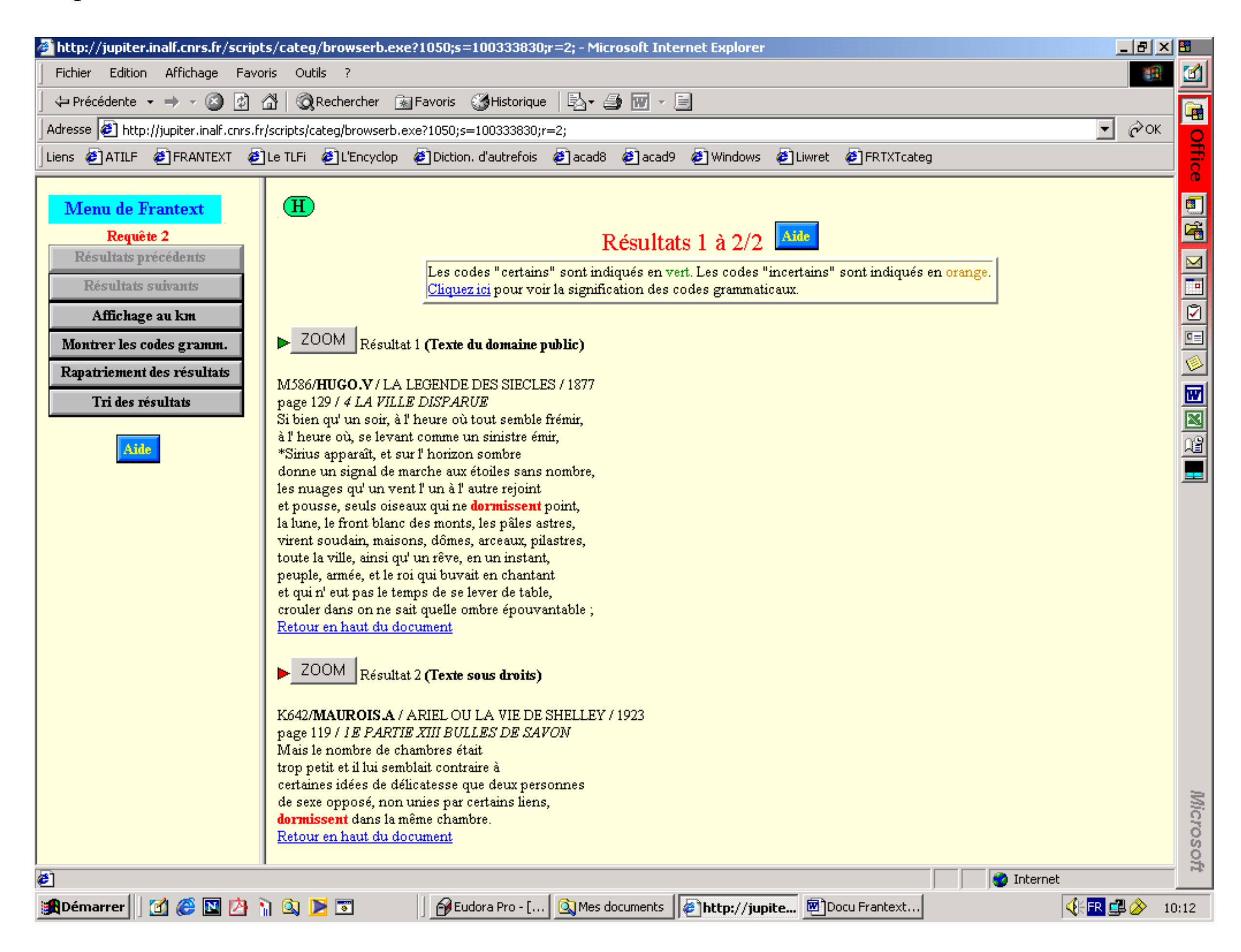
3. Résumé

- **Eléments pouvant entrer dans la composition d'une expression de séquence :**
  - **Une graphie donnée**
  - **&caimer** (forme conjuguée de aimer)
  - **&mcheval** (forme fléchie de substantif/adjectif ou participes d'un verbe)
  - **($Choix_1$ | $Choix_2$ | ... | $Choix_n$)** (avec possibité d'imbriquer à volonté)
  - **&?** (indicateur d'expression optionnelle. Ex. **un &?grand homme** ou **un &?(très grand) homme**)
  - **&e(XXX)** (entité catégorisée) avec possibilités pour XXX
    - **g=YY** ou **g!=YY** (catégorie grammaticale voulue/exclue)
    - **c=YY** ou **c!=YY** (contenu textuel voulu ou exclu)
  - **une invocation de règle de grammaire** (voir ci-dessous)
  - **le symbole de négation ^**
- **Les grammaires :**
  - Une grammaire est une suite de règles
  - Une règle est constituée d'un **nom** et d'un **corps**.
  - Le nom d'une règle est une chaîne de caractères composée des caractères de a à z (majuscule ou minuscule), de chiffres de 0 à 9, et des caractères - (moins) et _ (tiret de soulignement). Il doit être suivi du caractère : , puis d'un retour à la ligne.
  - Le corps de la règle peut s'étendre sur plusieurs lignes : un retour à la ligne dans le corps d'une règle n'a aucune signification.
  - Invocation d'une règle de grammaire :
    - **&rXXX** (invocation de la règle XXX de la grammaire courante)
    - **&rXXX,GGG** (invocation de la règle XXX de la grammaire GGG)
    - **&rXXX(P_1,P_2,...,P_n)** (invocation de la règle XXX de la grammaire courante avec passage de paramètres)
    - **&rXXX(P_1,P_2,...,P_n),GGG** (invocation de la règle XXX de la grammaire GGG avec passage de paramètres)
  - Les règles peuvent apparaître dans n'importe quel ordre. En particulier, le corps d'une règle peut invoquer une autre règle qui ne sera écrite que plus loin dans la grammaire.
  - **&1**,**&2**,**&3**,... manière générique de désigner, dans une règle de grammaire, le premier, le second, le troisième,... paramètre

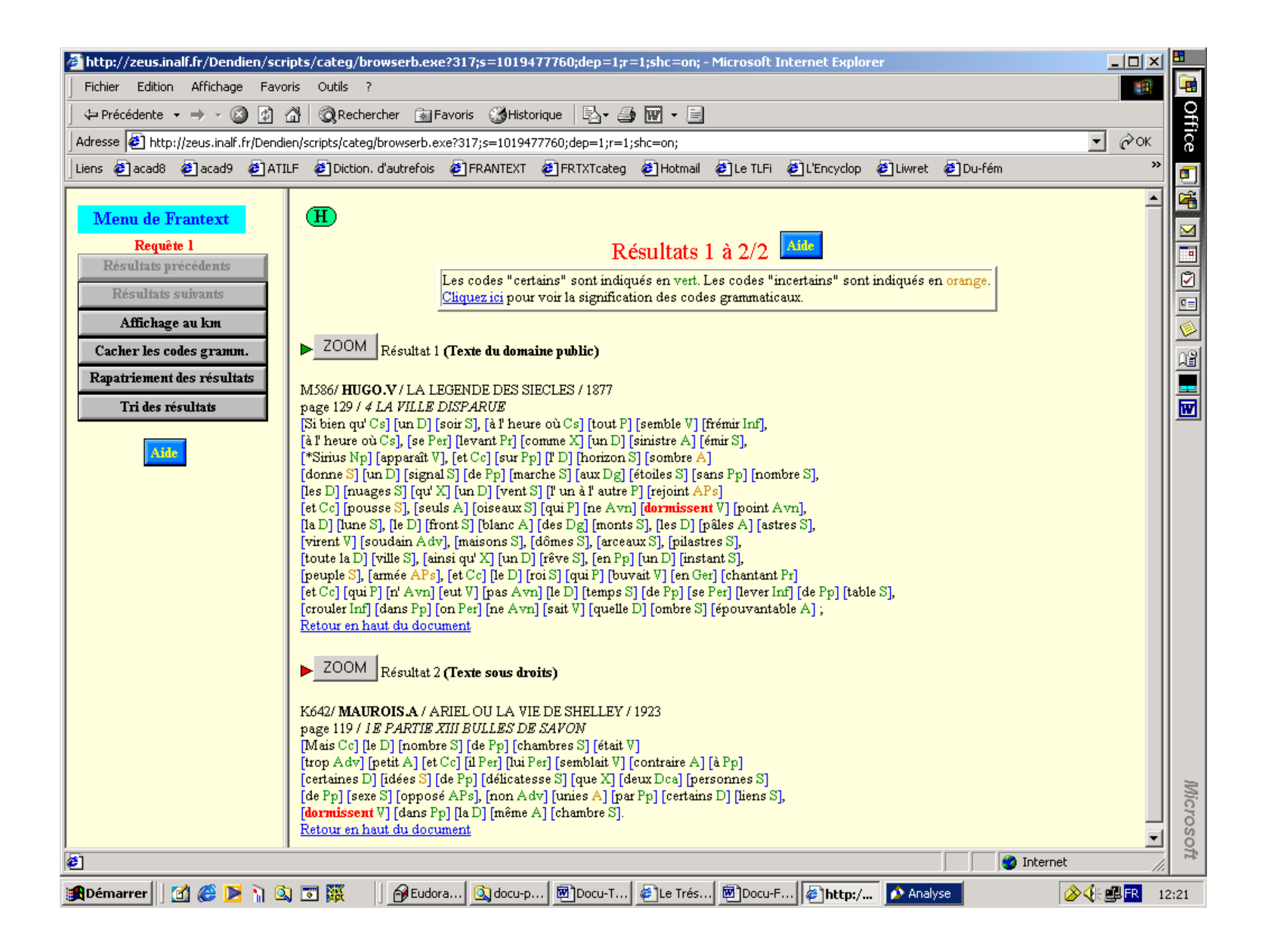
Exemples de requêtes

Exemple 1.

Corpus sélectionné = totalité de la base, soit 1940 textes
Séquence à rechercher = « *dormissent* »



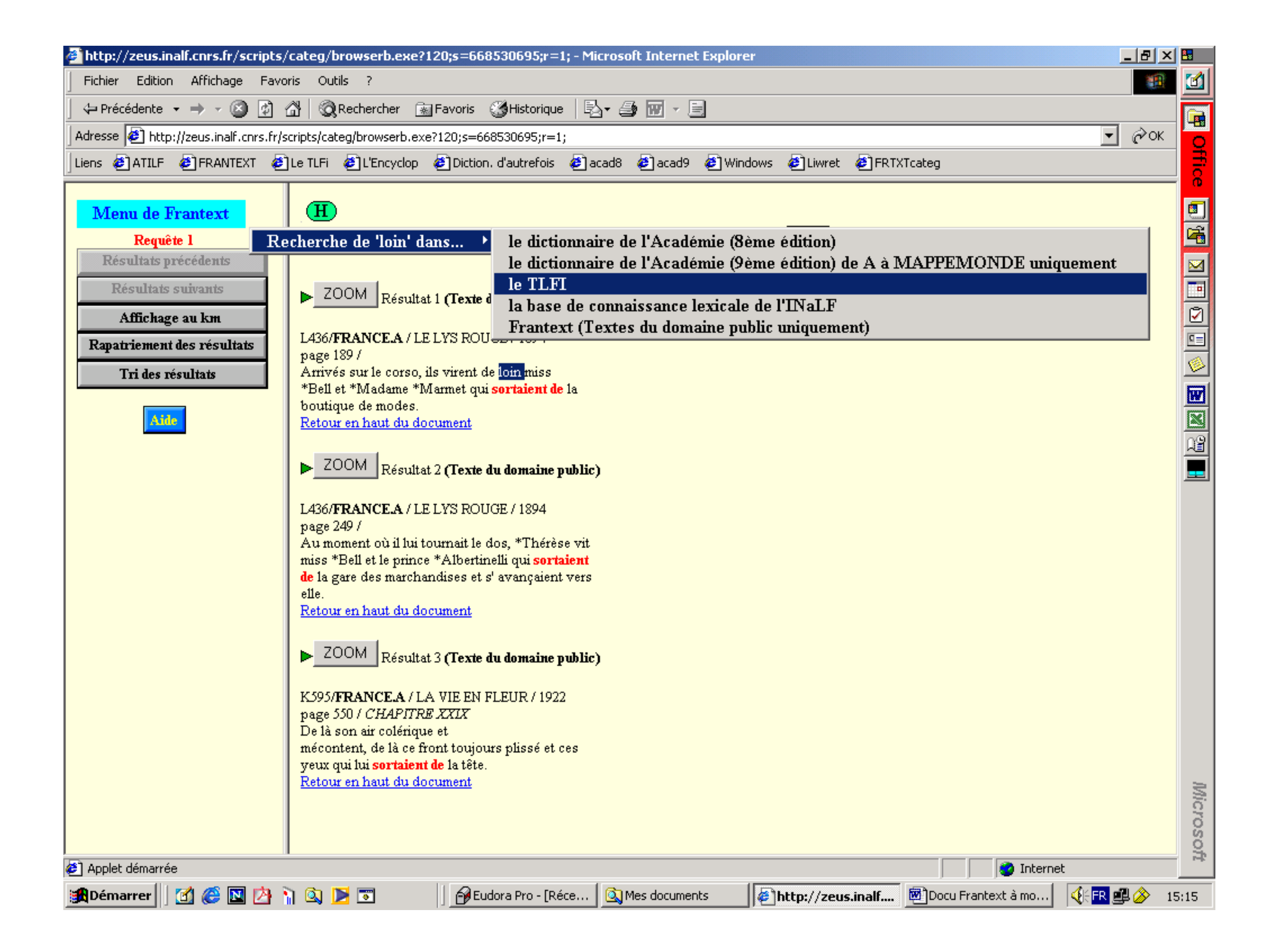
Propositions d'affichage au kilomètre, comme dans la base non catégorisée
de rapatrier les résultats,
de trier les résultats
mais aussi d'afficher les codes grammaticaux :



Exemple 2 :

Base catégorisée ou non catégorisée.
Corpus sélectionné par auteur = **France** -> 8 textes
Séquence à chercher = *sortaient de* -> 3 résultats

Si j'utilise l'hypernavigation, à partir du mot sélectionné = *loin*, le résultat s'affiche comme suit :



Exemple 3 :

Requête sur la base catégorisée ou non catégorisée.

3.1. Je définis le corpus : auteur = France

J'écris une grammaire simple, que j'appelle **jo**, et qui contient une seule règle, pour obtenir toutes les séquences *(goutte, marc, ration, tasse, gorgée, goutte)de café*.

http://zeus.inalf.cnrs.fr/scripts/categ/showp.exe?381;s=668530695;p=creat_gram.txt - Microsoft Internet Explorer

Fichier Edition Affichage Favoris Outils ?

Adresse http://zeus.inalf.cnrs.fr/scripts/categ/showp.exe?381;s=668530695;p=creat_gram.txt

Liens ATILF FRANTEXT Le TLFi L'Encyclop Diction. d'autrefois acad8 acad9 Windows Livret FRTXTcateg

Menu de Frantext

Création d'une grammaire

Note : Avant de créer une grammaire, il vous est conseillé de prendre connaissance du document Recherche dans les textes dans lequel sont présentées les différentes expressions utilisables dans les grammaires.

```
tousles :
(grain|marc|ration|tasse|gorgée|goutte)de café
```

N'oubliez pas d'indiquer ci-contre le nom de la grammaire à créer avant de cliquer le bouton "SAUVEGARDER" : jo SAUVEGARDER

Terminé Internet

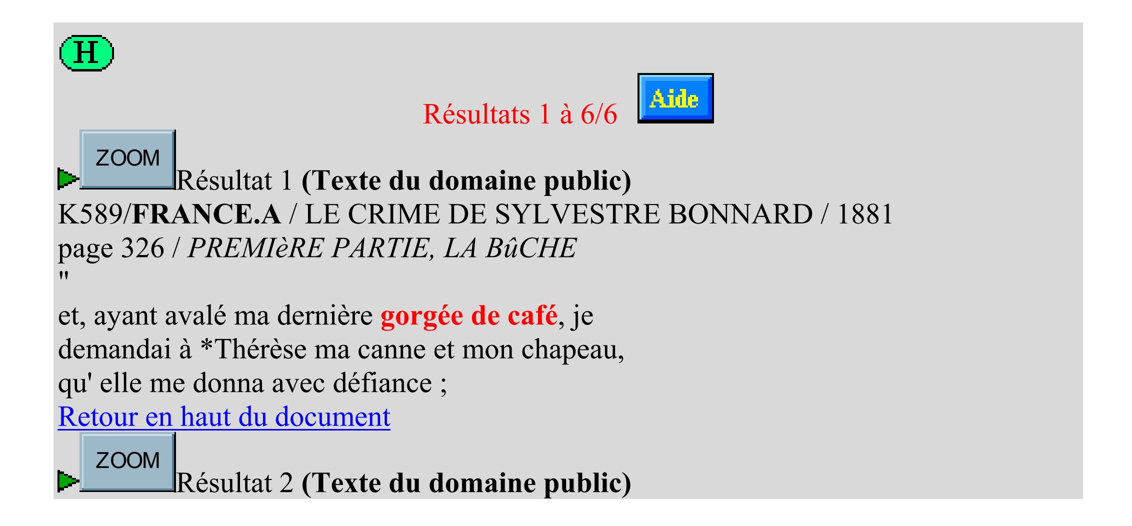
Démarrer Eudora Pro - [peter.blum... Mes documents http://zeus.inalf.cnrs.f... 15:53

J'invoque cette règle, et la grammaire qui la contient, dans le formulaire de requête (Lancer une recherche) :

Aide **1) Partie du formulaire à remplir obligatoirement**

Définition de la séquence 1 : &rtousles,jo

Et j'obtiens 6 résultats :



H

Résultats 1 à 6/6 Aide

ZOOM Résultat 1 **(Texte du domaine public)**

K589/**FRANCE.A** / LE CRIME DE SYLVESTRE BONNARD / 1881

page 326 / *PREMIèRE PARTIE, LA BûCHE*

"

et, ayant avalé ma dernière **gorgée de café**, je
demandai à *Thérèse ma canne et mon chapeau,
qu' elle me donna avec défiance ;

Retour en haut du document

ZOOM Résultat 2 **(Texte du domaine public)**

L436/**FRANCE.A** / LE LYS ROUGE / 1894
page 384 /
Une **tasse de café**,
presque vide, était sur la table.
Retour en haut du document

ZOOM Résultat 3 **(Texte du domaine public)**
K724/**FRANCE.A** / L'ILE DES PINGOUINS / 1908
page 414 / *L. 8 TEMPS FUT., HIST. SS FIN*
Les
soldats firent avec entrain le service d' ordre et
reçurent une double **ration de café**.
Retour en haut du document

ZOOM Résultat 4 **(Texte du domaine public)**
K592/**FRANCE.A** / LES DIEUX ONT SOIF / 1912
page 236 / *CHAPITRE XIX*
Toute la journée, il griffonnait sur ses genoux, au
pied de son lit, trempant des tronçons de plumes
usées jusqu' aux barbes dans l' encre, dans la suie,
dans le **marc de café**, couvrant d' une illisible écriture
papiers à chandelle, papiers d' emballage, journaux,
gardes de livres, vieilles lettres, vieilles factures,
cartes à jouer, et songeant à y employer sa chemise
après l' avoir passée à l' amidon.
Retour en haut du document

ZOOM Résultat 5 **(Texte du domaine public)**
K594/**FRANCE.A** / LE PETIT PIERRE / 1918
page 52 / *CHAPITRE VIII*
Je ne sortais jamais de chez *Corcelet sans avoir
pris un **grain de café** que je mâchais en chemin.
Retour en haut du document

ZOOM Résultat 6 **(Texte du domaine public)**
K594/**FRANCE.A** / LE PETIT PIERRE / 1918
page 157 / *CHAPITRE XXI*
Mon parrain, sa **tasse de café**
à la main, me rejoignit à la fenêtre et me
demanda où était le papegai.
Retour en haut du document

3.2. Une grammaire jo2 écrite différemment, avec la même séquence, mais précédée du symbole de négation,

saufles :
^(grain|marc|ration|tasse|gorgée|goutte)de café

amène 5 résultats (ces séquences sont exclues)

3.3.Une grammaire jo4 contenant :

tousles :
(&mgrain|&mmarc|&mration|&mtasse|&mgorgée|&mgoutte|&mservice|&mmoulin)(de|à) café

appelle 11 réponses (les séquences sont au singulier ou au pluriel).

3.4. Une grammaire jo5 contenant :

saufles :
^(&mgrain|&mmarc|&mration|&mtasse|&mgorgée|&mgoutte|&mservice|&mmoulin)(de|à) café

appelle3 réponses (toutes ces séquences sont exclues).

Exemple 4

Sur la base catégorisée uniquement
Une grammaire **jo6** contenant une règle :

sans:
&e(g=V c=(&capporter|&cvouloir)) &q café

amène 23 résultats :

Résultat 1 **(Texte sous droits)**
M325/**DUMAS.A PERE** / LE COMTE DE MONTE-CRISTO / 1846
page 407 /
*Ali **apporta le café**.

Résultat 2 **(Texte sous droits)**
M734/**FLAUBERT.G** / MADAME BOVARY / 1857
page 22 / *I CHAPITRE 3*
On **apporta le café** ;

Résultat 3 **(Texte sous droits)**
L829/**PONSON DU TERRAIL** / ROCAMBOLE T.5 / 1859
page 60 / *EXPLOITS DE ROCAMBOLE 2*
à la fin du dîner, le valet qui les servait à table
apporta du café et des liqueurs.

Résultats suivants :

capitaine d' habillement, quand on **apportait le café**
piano, pendant qu' on **apportait le café**, et jouai
*Mme *Faujas descendait, lorsque la cuisinière **apporta le café**.
-vous savez, si vous **voulez du café**, il y en a là.
pauvre, tombée au rôle de domestique, lui **apportait son café**.
Ensuite, il **apporta le café** ;
on **apportait le café**, la gaieté fut encore accrue
On nous **apporte du café**.
-et j' **apporté le café**, fit *Mme *Bavoil ;
On **apportait le café**.
*Wazemmes **apporta le café**.
On **apportait le café**.

qui lui **apportait du café** au lait et des brioches.
*Magnin **voulait du café**.
Au moment où *Mathurine **apportait le café**, le
des deux bonnes, lui **apportait son café** au lit.
-je **voudrais un café**.
tu nous **apporterais le café** à Annie et à moi.
Puis avec décision: j' **apporte le café**.
Laurent lui **apporte du café** puis, quand ils ont bu,

Table des matières :



ATILF (UMR 7118 CNRS - Université Nancy 2)
44, avenue de la Libération - BP 30687 - 54063 Nancy Cedex
Tél. 03 83 96 21 76 Fax 03 83 97 24 56
Site : www.atilf.fr Courriel : contact@atilf.fr



